Ancient Genomes link early farmers from Atapuerca in Spain to modern-day Basques

Supplementary Information

Torsten Günther*, Cristina Valdiosera*, Helena Malmström, Irene Ureña, Ricardo Rodriguez-Varela, Óddny Sverrisdóttir, Evangelia A. Daskalaki, Pontus Skoglund, Thijessen Naidoo, Emma M. Svensson, José María Bermúdez de Castro, Eudald Carbonell, Michael Dunn, Jan Storå, Eneko Iriarte, Juan Luis Arsuaga, José Miguel Carretero, Anders Götherström, Mattias Jakobsson[†]

*These authors contributed equally
[†]correspondence to: mattias.jakobsson@ebc.uu.se

Table of Contents

S1. The El Portalón site stratigraphic and cultural sequence

The El Portalón cave is an archaeological site in the Sierra de Atapuerca (Burgos, Spain; Fig. S1), a region well known for its Early and Middle Pleistocene sites [1,2], but more recent excavations have shown that it also presents a rich and varied archaeological record [3–5]. Previous studies have been successful in obtaining ancient DNA from faunal remains in the El Portalón cave [6,7].

The El Portalón cave spans the Late Pleistocene to the Holocene and it is during this later period that the cave became one of the main entrances to the Cueva Mayor-Cueva del Silo karst (Fig. S1). Recent excavations at the El Portalón have revealed a stratigraphic sequence starting in the Late Pleistocene and showing evidence of human occupation throughout the Holocene. [3] reported detailed radiocarbon dates for the entire stratigraphy ranging from 30,000 BP to 1,000 BP. Two major sedimentary phases are observed in the El Portalón archaeological record, the lower phase comprises sediments from the Late Pleistocene and has a significant paleontological record and sparse Upper Palaeolithic human artifacts. The upper phase corresponds to the Holocene and is characterized by dark archaeological sediments with abundant archaeological artifacts. The Holocene phase includes Mesolithic, Neolithic, Chalcolithic, Bronze Age, Iron Age, Roman and Medieval periods of human occupation. The palaeoenvironmental data obtained from the speleothem record [8], are of great importance because of the scarcity of this type of information regarding the Upper Palaeolithic, Mesolithic and Neolithic cultural periods in the interior of the Iberian Peninsula. In addition to its vast cultural/technological and faunal remains, a series of burials have been identified in the Chalcolithic period. These burials are often accompanied with grave goods, such as pottery and small animal bones in anatomical position, most commonly sheep. The archaeological content combined with direct radiocarbon dating of the human samples analysed suggest a pre-Bell Beaker Chalcolithic chronocultural assignation for the burials.

S1.1. Early Chalcolithic (Pre-Bell Beaker) funerary context, Bronze Age and disturbed layers from the El Portalón cave

The burial phase comprises a tumular stacking of decimetric limestone clasts (Figs. S1 & S2). The tumulus appears to be of an oval shape and its structure corresponds to a progressive accumulation of limestone clasts in an agrading (vertical) and prograding fashion over a basal surface defined by floors that have ceramic "pavements" on which there are numerous circular small fire pits filled with partially combusted charcoal fragments. Associated with the fire pits, are nearly complete small ceramic bowls, along with remains of juvenile individuals of domestic fauna, mainly lambs, in anatomical connection (Fig. S2). Among the limestone clasts and sometimes lying on the floor, are different human remains, defining funerary contexts with some archaeological elements typical of grave goods (Fig. S2).

For example, within the funerary contexts, there is a large amount of archaeological material associated to the human remains. Pottery fragments are the most abundant archaeological remains (50%) and although very fragmented, most of them are simple forms, of which closed, globular and ovoid shapes prevail. Ornamental decoration is simple and homogeneous, the motifs are reduced to just a few themes such as cylindrical perforations, impressions, embossed tablets, simple incised or channelled rope lines or incisions. In general, most of these decorations are located below the rim (Fig. S2). The pottery types and decorations, in particular the embossed tablets, are characteristic of the Early Chalcolithic (Pre-Bell Beaker) phase and common for the Chalcolithic in the Duero Basin [9].

The bone industry collection is composed of nineteen pieces that can be divided into two functional groups: bone and shaft tools and personal adornment objects. Among the group of "tools", awls, a gradine, a bipoint, a rod and a spatula with ochre residue have been identified. Regarding the lithic materials, 260 pieces (3%) associated to the tumulus have been recovered and with the exception of two polished elements, the rest are from the carving industry (Fig. S2). The number of adornment objects associated with the funerary context is very scarce, as only seven remains have been recovered.

Within the tumulus, 3,694 faunal bone remains have been recovered, of which 51.2% (1,892) have been anatomically and taxonomically identified and the other 48.8% (1,802) are of undetermined remains (bone fragments or shards). From this large faunal collection, ovicaprines are the most abundant (*Ovis aries* and *Capra hircus*) (41.65%), followed by *Bos taurus* (11.15%), *Sus domesticus* (7.82%), *Canis familiaris* (1.64%) and *Equus* sp. (0.9%). Hunting activity is indicated by the presence of *Cervus elaphus, Capreolus capreolus, Vulpes vulpes, Leporidae indet* and a small carnivore. Furthermore, a small percentage of fish and turtle shells have been identified.

We suggest that the El Portalón pre-Bell Beaker Copper Age funerary barrow is the result of repetitive burial activities following a similar pattern over time. Each event appears to have disturbed a previous funerary context. In addition, the collapse of a large part of the roof of the cavity during this period and its later use as a habitat and animal stable seemed to have contributed to the disturbance of some of the funerary structures. Under such events, human bone remains are a common occurrence, and grave goods elements often appear scattered among the limestone blocks of the tumular structure (Fig. S2).

Chalcolithic funerary practices in El Portalón were prolonged over six generations (ca. 150 yr), creating a collective but not simultaneous burial site. The continuous use of this space has altered the primary burials, displacing and scattering the bones; usually only isolated bones or small bone sets remain at different levels of anatomical connection.

In this tumular funerary structure, we have discerned human remains from a minimum of eight individuals (three adults and five subadults). In just one case, we have found a practically intact burial and the skeleton of the individual with grave goods (Fig. S2). The rest are isolated and partially articulated bones found in secondary position. The well-preserved burial, provides significant information regarding the potential characteristics of the funeral rites conducted during the Chalcolithic period at this cave (Fig. S2). This burial belongs to a complete subadult individual found in primary position accompanied by grave goods.

According to [10], this individual corresponds to a male child of an estimated height of 101 cm and due to the state of tooth eruption, he must have died at around 6-7 years of age. Macroscopic analysis and computerized axial tomography reveal a set of lesions both on the cranium and on the long bones, indicating that this individual likely suffered from rickets and/or scurvy at different stages of his life. The etiology of the two metabolic diseases could be attributed to an abnormal diet [10].

The boy´s burial was accompanied by different grave goods; for example, his feet, legs, pelvic area and head were covered by pieces of large fragmented ceramic (Fig. S2). On top of the body and the ceramic, a covering of green clay was deposited. Comprising part of the ensemble and also covered by the same clay, a nearly whole calf skeleton (*Bos taurus*) was found in anatomical connection. There are substantial amounts of pottery fragments (more than 200), including all vessel parts (rims, bodies, bases) among the grave goods associated with this burial. These fragments have allowed for the reconstruction of several vessels, among which a truncated cone-shaped morphology is notable (Fig. S2).

In addition to the pottery, several tools were found as part of the boy´s grave goods, among these, the most notable ones were a pedunculated arrowhead, a medial fragment of bitruncated white flint flake with marks from wear on both blades, a simple honey-colored flint flake of color with retouching on one of its blades, a quartzite un-retouched spall and a tubular awl on the left distal tibia of the ovicaprine with beveling on the far end, showing cut marks and abrasion prints.

The presence of other burials in the excavated area of the tumular structure is reflected by the existence of sets of human bones (in some cases in anatomical connection). The minimum number of individuals represented by the bones associated to the funerary tumulus (211) is eight: three adults and five subadults including the aforementioned complete child (6-7 years of age). This makes a total of 8 individuals that have been found within the funerary tumulus. Note that while we report aDNA recovery from 8 individuals only 5 come from the tumulus, the others come from the Bronze Age level (n=1) and the clandestine excavation (n=2), detailed below.

Furthermore, 43 additional human bone fragments have been recovered in Middle Bronze Age levels at the space known as the Salón del Coro or Galería Principal, which is also part of the El Portalón site itself.
Finally, in the 2001 field season we identified an area of clandestine excavation in the central part of Portalón carried out by unknown individuals [3]. From 2001 to 2006 we excavated this disturbed area in which we founded significant archaeological and paleontological materials out of their original context, including 91 human bones.

S1.2. Sample provenance and radiocarbon dating

Sixteen human samples were selected from a series of human remains recovered during our yearly field excavations conducted since the year 2000 to 2012, but only eight of these are reported in this study (see Table S1). Most samples (n=5) correspond to individuals found in the Chalcolithic funerary tumulus or Bronze Age levels (n=1), whereas the remaining two were recovered from the clandestine pit (see above). However, all samples were directly radiocarbon dated to the Chalcolithic and Bronze Age (Table S1). The repository of all bone and tooth remains is the Laboratory of Human Evolution at the University of Burgos (Spain).

The sixteen human samples were radiocarbon dated using accelerator mass spectrometry (AMS) at Beta Analytic Inc. (Miami, Florida). Calibration to years cal BP was made using Oxcal v4.2.3 software based on the IntCal13 atmospheric curve [11] (Table S1). From these sixteen individuals, only eight were successfully sequenced (i.e. at least one library showed more than 1% human DNA) and only data from these 8 individuals are shown (Table S1).

S2. Sample preparation, DNA extraction, library construction and sequencing

We obtained genomic sequence data from bone and tooth samples belonging to 16 ancient farmers. All samples were prepared in dedicated ancient DNA (aDNA) facilities at the Evolutionary Biology Center in Uppsala, Sweden.

S2.1. Bone and teeth DNA extraction

The first millimeter of the bones and teeth was abraded using a Dremel™ tool and then ground into powder using a multitool drill (Dremel™). Approximately 250mg of this bone/tooth powder was used for DNA extraction following three silica-binding methods ([12–14] with modifications as in [15]). We made between 2 and 5 DNA extractions for each sample. All extractions were done in batches of eight including 7 samples and one extraction blank.

S2.2. Sequencing Library Building

Given the degraded nature of ancient DNA (highly fragmented), we constructed DNA libraries by skipping the initial nebulization step.

Twenty microliters of extracted DNA were converted into Illumina multiplex sequencing libraries (blunt end ligation method), following [16]. One to two libraries per extract were built resulting in a total of 2-6 libraries per individual. DNA libraries were enriched by amplifying 6 PCR reactions for each library. Library amplification was carried out in a final volume of 25µl using AmpliTaq Gold® DNA Polymerase (Life Technologies) with a final concentration of 1X Gold Buffer, 2.5 mM Magnesium Chloride, 250µM dNTP (each), 3µl of DNA library, 0.2µM IS4 PCR primer (5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT 3’) and 0.2µM indexing primer (5’-CAAGCAGAAGACGGCATACGAGATxxxxxxxGTGACTGGAGTTCA GACGTGT, where x is one of 228 different 7bp indexes provided in [16] and 0.1 U/µl of AmpliTaq Gold. Two negative controls were added in each PCR. Cycling conditions were performed as follows: a 12 min activation step at 94 ºC, followed by 8-15 cycles of 30 s at 94 ºC, 30 s at 60 ºC, 45 s at 72 ºC, with a final extension of 10 min at 72 ºC. The number of cycles was different for each sample and varied between 8 and 15. All 6 PCR reactions were pooled and purified using the AMPure XP (Agencourt-Beckman Coulter A63881) following the manufacturer’s guidelines. After amplification, the concentration and size profiles of the purified libraries were determined on a Bioanalyzer 2100 using the High Sensitivity DNA chip (Agilent) for DNA visualization. None of the extraction or PCR blanks showed signals of DNA on agarose gels and Bioanalyzers and were therefore not furthered sequenced.

Finally, all libraries were pooled at equimolar concentrations and shotgun sequenced (100bp, paired end) on Illumina’s HiSeq2000 platform at the SNP & SEQ technology platform SciLife Sequencing Centre in Uppsala. The initial sequencing of libraries was done for screening purposes, where we would select our best libraries based on human DNA content and library complexity for further sequencing. Libraries that showed less than 1% of total human DNA were discarded for further sequence. As a result, only eight out of the initial 16 individuals were used for downstream analyses. Our sequencing effort is summarized in Dataset S8.

S3. Sequence processing and alignment

Read pairs were merged simultaneously as the adapters were removed using, requiring an overlap of at least 11 bp and summing up base qualities using MergeReadsFastQ_cc.py [17]. Merged reads were mapped as single-end reads to the human reference genome (build 36 and 37) using *BWA aln* version 0.7.8 [18] with the non-default parameters -n 0.01 -o 2 and disabled seeding as in [19,20]. PCR duplicate reads with identical start and end coordinates were collapsed into consensus sequences using FilterUniqueSAMCons.py [17]. The effect of potential mapping errors is further minimized by just using known variants (see Supplementary Material S7). We also remapped ancient comparative data (see Supplementary Material S7) using the same procedure. We required less than 10% mismatches to the human reference genome for each read, and discarded reads with a length of less than 35 bp. The sequences showed deamination pattern towards fragment-ends which are characteristic for ancient DNA [21] (Figure S3).

S3.1. Mitochondrial DNA haplogroups

Consensus sequences for the mitochondrial genomes of all samples were called using *mpileup* and *vcfutils.pl* (*vcf2fq*) from the *samtools* package [22]. Only reads with a minimum mapping score of 30, a minimum base quality of 30 and a minimum coverage of 3x were used to call confident bases for these consensus sequences. We then used *haplofind* [23] and PhyloTree Build 16 [24] to assign sequences to known haplogroups.

S3.2. Mitochondrial DNA authentication

We estimated mtDNA contamination by identifying private or near-private (<5% in 311 modern mtDNAs) consensus alleles in each ancient individual [25]. We retained reads with a minimum mapping score of 30 and positions where coverage was at least 10 and base quality was 30. We excluded positions where the consensus allele was the C or G in a transition polymorphism, as these are vulnerable to post-mortem misincorporations. A point estimate *c* of mtDNA contamination was estimated by adding the counts of the consensus and the alternative bases across all sites, assuming independence among positions, where $\hat{c}=N_{\text{alternative}}/(N_{\text{consensus}}+N_{\text{alternative}})$. If no alternative allele was found, we estimated an upper confidence limit on contamination using a binomial distribution

$$P=\binom{N}{k}c^{k}(1-c)^{N-k}$$

where N=$N_{\text{consensus}}$+$N_{\text{alternative}}$, and k=0. We then found the value of c where P=0.05. When alternative alleles were observed, a 95% confidence interval was computed using a binomial approximation

$$95\,\mathrm{CI}=\mathrm{c}\pm 1.96\sqrt{\frac{c(1-c)}{N}}$$

Furthermore, we used a Bayesian approach which checks whether mitochondrial reads map better to the consensus sequence or a reference data set of 311 modern human mtDNAs [26]. Contamination estimates for high coverage mt genomes were consistent with the results of the method described above (Table S2).

S3.3. X chromosome based authentication

Since males carry only a single copy of the X chromosome, it can also be used to estimate contamination. We applied the method introduced in [27] to all male samples in our analysis. The approach utilizes the fact that contamination would only cause discordant bases at polymorphic sites while sequencing and mapping error will affect non-polymorphic sites as well. So by comparing mismatches at SNP sites to the adjacent non-polymorphic sites, an estimate for contamination can be obtained [27].

Polymorphic transversion sites were obtained from the Iberian population (IBS) of the 1000 genomes project phase 3 [28] using vcftools [29]. Since this population is also a likely source for contamination, we also estimated allele frequencies from the IBS samples. We excluded regions of the X chromosome that are homologous to the Y chromosome as well as regions filtered for mappability (100mer, [30,31]) to avoid ambiguous read mapping. Only reads with a mapping quality score of 30 or more and bases with a quality of at least 30 were included in the analysis. Due to the low coverage nature of most samples, we did not filter for read depth. ANGSD [32] and the R scripts provided with the package were used to apply this method. Results using all reads (Method 1) and a single read per site (Method 2) are shown in Table S3.

S3.4. Biological sexing

The biological sex of all individuals was assessed using the ratio of reads mapping to the X and Y chromosomes (R_y) which is then compared to a reference panel [33]. We restricted the analysis to sequence alignments with mapping quality of at least 30.

S4. Mitochondrial analysis

The mitochondrial genome coverage varied between 3.6x and 341x among the 8 individuals. The number of SNPs that supports each called haplotype varied between 41 and 50 and they are reported as deviations from the Reconstructed Sapiens Reference Sequence (RSRS) [34] (Table S4). Six of the consensus sequences lacked sequence data at between one and seven of the hg defining position for the called haplotype. A few additional mutations that can be found elsewhere in the mtDNA phylogeny were observed in the sequences. These were mainly transitions at positions with low coverage, and even though a few of these may be true mutations, most can probably be attributed to post-mortem deaminations. Note that polymorphisms reported at np 195, 4769, 8860, 14766 and 16519 are derived in relation to RSRS but considered ancestral in relation to rCRS [35]

All eight individuals from Atapuerca displayed unique haplotypes (Figure 1B, Table S2, Table S4). The most abundant haplogroup, U5, was found in three temporally non-overlapping individuals. Two belonged to subtypes of U5b (U5b3 and U5b1b) and one belonged to U5a (U5a1c). Note that the U5b1b individual (ATP9) have the T6189C back-mutation but display the ancestral state for G12618A. Two individuals belonged to H3. They were dated to within the same time-frame but were not maternally related as one of them carried a T to C transition at np 12957 classifying it to H3c. The remaining three individuals belonged to the haplogroups J, K and X (J1c1b1, K1a2b and X2c).

The mitochondrial lineages of the ATP individuals show a heterogeneous ancestry and can be traced back both to hunter-gatherer (HG) and subsequent farmer contexts (Dataset S2 (external file), Figure S4). The most frequent haplogroup in ATP, U5, is commonly found in HG groups in Iberia and across Europe and Scandinavia [19,20,36–43]. U5 subhaplogroups are also found in Neolithic farmer populations in Europe although at lower frequencies [36,42,44–48]. The remaining four haplogroups found in ATP, H, J, K and X, are present in other farmer populations from the Neolithic and onwards [20,38,48–50]. In southern Europe (e.g. Spain, Portugal and Italy), however, haplogroup H is also frequent in Paleolithic and Mesolithic HG populations [26,36,42].

Even though some haplogroups (U5b and H) are shared between ATP and HGs from Mesolithic Iberia (southern hunter-gatherers SHG), the general haplogroup composition between the groups differ [19,36,43] (Dataset S2, Figure S4), similar to the differences between other farmer and HG populations in Europe[19,20,37–40,48]. None of the previously investigated Neolithic farmer populations from Iberia have similar haplogroup distribution as ATP (Dataset S2, Figure S4). These farmer groups also differ from each other [42,51]. Analysis of haplogroup frequency data have for example shown that early Neolithic north-eastern Iberian populations cluster with early- and middle Neolithic populations from central Europe while other Neolithic Iberian populations (from Basque Country and Navarre, NBQ [42] and Portugal, NPO [36]) share a closer affinity to HG populations [48]. NBQ is the population that share the largest number of haplogroups with ATP (X, H, J, U5b and K [42] although the frequencies differ and NBQ also display additional haplogroups (U, T, HV and I). The Chalcolithic individuals from El Mirador (MIR), a cave located in the same mountain system as ATP (Sierra de Atapuerca), present a somewhat different haplogroup distribution than ATP. MIR clusters with early Neolithic Iberians and early and middle Neolithic central European populations [52]. They lack the U5 subhaplogroups found in ATP and instead display T2 and U3 [52]. RFLP data from another Chalcolithic population from the Basque

Country show the same main haplogroups as found in ATP and MIR (38% H, 17% U, 13% J, 21% K and 9% T+X, note that these frequencies are not added to Dataset S2 or to Figure S4 as they are not generated from sequence data), although the lower resolution of the data cannot specify which population (ATP or MIR) that it is most similar to [53]. It has further been suggested that the mt-haplogroup composition of Basque populations differs between Chalcolithic [53] and historical times (600-700 AD) [54] with increasing frequencies of H anv V haplotypes and with increasing similarities to present-day western European populations. The picture of the ancient farmers in Iberia remains unresolved and the limited level of information retrieved from mitochondrial DNA has not been able to go beyond the above described observations.

Present-day European populations are genetically quite homogenous in terms of mitochondrial haplogroup distributions and it is mainly haplogroup frequency differences that separate different populations. It is therefore not a straightforward process to assess the potential connections between ATP and specific present-day populations. We note that the most abundant lineages in the ATP individuals are found in higher frequencies in some Basque-speaking populations (U5b; [55], and H3; [55,56]) than in other European populations. Further, several haplotypes have been suggested to be autochthonous to present-day Basque populations (see e.g. [57–60]). Two of these are J1c1 [61], a lineage ancestral to the J2c1b1 haplotype in ATP7, and H3c2a [57], a lineage that derives from the H3 and H3c haplotypes found in ATP17 and ATP12-1420 .

S5. Y chromosome analysis

A list of informative SNP sites was obtained from AMY-tree v2.1 [62], using a snapshot of PhylotreeY [63]. We excluded all non-SNP sites, transition sites (to avoid deamination damage), and A/T and G/C SNPs (to avoid strand misidentification). This left us with 550 sites after filtering. All these sites were checked in ATP12-1420 and ATP2 for bases with a minimum quality of 30 and a minimum mapping quality of 30. We also excluded sites showing two different alleles since those are likely sequencing errors or contamination, this excluded 1 SNP for ATP12-1420 and 14 SNPs for ATP2.

S5.1. ATP2

ATP2 displayed the derived allele for nine Y chromosome markers (Table S5); with all of the markers providing phylogenetic support for ATP2 belonging to haplogroup H2. These markers included: L985, L1013 and L1053 (A1); M235, P159 and P187 (F); L279, L281 and P96 (H2). Previously labeled haplogroup F3, H2 was recently redefined on the basis of an overlap between the datasets of [64] and [65] [63]. It was found that the two haplogroups, H-M69 and F3-M282, shared a root defined by the marker M3035. While only a few H2 individuals have ever been found, the haplogroup appears to have a west Eurasian distribution; with a low level Middle Eastern presence in modern-day Iran, Turkey, Bahrain, Kuwait and Qatar (Family Tree DNA, [66]), as well as minor occurrences in modern-day England, France, Sardinia, Sweden and the Netherlands (Family Tree DNA, [64,65,67]). H2 also seems to occur at low frequencies in Neolithic samples [68].

S5.2. ATP12-1420

ATP12-1420 displayed the derived allele for five Y chromosome markers (Table S6); with all of the markers providing phylogenetic support for ATP12-1420 belonging to haplogroup I2a2a. These markers included: L1053 (A1); P124 (IJ); L460 (I2a); L34 and P221 (I2a2a). While the almost European-specific haplogroup I arose approximately 20000 to 25000 years ago [67,69], haplogroup I2a2a may have diverged as a subclade, around 15000 years ago [69], possibly during the recolonization of Europe following the Last Glacial Maximum (LGM). Unlike the more common subclades of I1 and I2a1, haplogroup I2a2a appears at relatively low frequencies across much of Europe. Its highest levels (10-12%) are found in modern-day Germany and the Netherlands, with frequencies of around 5%, notably occurring in parts of modern-day France as well as Mordvin in the Volga region of central Eastern Europe [69].

Other members of haplogroup I have been discovered previously in ancient individuals; e.g. I* in Mesolithic Scandinavians [19], I1 in Hungary [70], I2a in Neolithic individuals from Hungary [71] and France [46], I2a1 in Neolithic Croatia [70] and a late hunter-gatherer from Sweden [20], and I2a1b in Mesolithic individuals from Luxembourg and Sweden [19].

Other haplogroups found among ancient specimens include C* in Upper Paleolithic Russia [72] and Mesolithic Spain [73], C6 in Neolithic Hungary [71], E1b1b1 in Neolithic Spain [45], F* in Neolithic Germany [44] and Neolithic Hungary [70], G2 in Neolithic Hungary [70], G2a in Neolithic France [46], Neolithic Germany [44], Neolithic Hungary [70], Chalcolithic Italy [74], and Neolithic Spain [45], J2a1 in Bronze Age Hungary [71], K (xLT) in Upper Paleolithic western Siberia [75], N in Iron Age Hungary [71], R* in south central Upper Paleolithic Siberia [76], R1a in Neolithic Germany [77], and Neolithic R1b in Germany [78].

S6. Phenotypic/selection analysis

In order to get an idea about the appearance of our individuals and their allelic status at sites identified to evolve under selection in modern-day Europeans, we checked reads covering five specific SNPs and compare them to a Mesolithic Iberian [73]. We only included reads with a mapping score of at least 30 and sites with a minimum base quality of 30 into the analysis. A more comprehensive study of selection targets and phenotypes would require a larger sample size and higher coverage per individual especially since most of these phenotypes are highly polygenic. Furthermore, we note that the potential post-mortem damage at transition sites might affect the conclusion of such surveys. Allelic states are shown in Table S7.

As shown before [99], the inhabitants of the El Portalon cave were probably all lactose intolerant in adulthood. This suggests a much later spread of this variant that has been the target of adaptation to a milk-rich diet in modern-day northwestern Europeans which also occurs at reasonably high frequencies in Northern Spain and Basques [100]. All sequences for the SLC24A5 (rs1426654) variant showed the derived state in the El Portalon individuals, which together with two derived variants at SLC45A2 (rs16891982), suggest that the pigmentation of the Chalcolithic Iberians was lighter than the Mesolithic LaBrana1 individual who carried the ancestral states at these major pigmentation loci [73].

rs1805007 in MC1R which is associated with red hair and light skin is ancestral in all but one sequence (out of eleven in all El Portalon individuals) but that single derived base call might also be due to post-mortem damage. rs12913832, a SNP that explains more than 56% of the variation between blue and brown eyes [101], has been shown to be derived in the Mesolithic LaBrana1 [73]. Two of the El Portalon individuals show only ancestral alleles at this site whereas one individual shows both variants suggesting the individual is heterozygot at the site. These observations suggest some eye color variation but also a tendency towards brown eyes in the Chalcolithic Iberians.

To summarize, Chalcolithic Iberian farmers seem to be lactose intolerant as the Mesolithic inhabitants of the Peninsula. However, their pigmentation was fairer and their eyes were darker than in the hunter-gatherer LaBrana1.

S7. Comparative data from modern populations

S7.1. SNP genotype data

The ancient individuals were merged with the SNP genotype calls of the Human Origins data set [19,79]. For each site and ancient individual, we randomly picked one read with minimum mapping quality 30 covering that site and used the respective base as allele if its base quality was 30 or higher. Sites showing indels were excluded. We also excluded all transition sites to avoid potential post-mortem damage. Ancient individuals used in this study and the number of SNPs after merging with the human origins data set are shown in Table S8. To increase power, only El Portalon individuals with more than 20,000 transitions SNPs were used for all major analyses.

S7.2. Sequence data

VCF and BAM files for the Yoruban individuals sequenced in phase 3 of the 1000 genomes project [28] were obtained from ftp.1000genomes.ebi.ac.uk. We used vcftools [29] to extract all SNPs with a minor allele frequency of at least 10% in the Yoruban population and excluded all transition polymorphisms to avoid potential post-mortem damage. The remaining 1,938,919 SNPs were merged with the ancient genomes and Denisovans ([80], as an outgroup) as described above (Table S9).

S8. Population genetic analysis

S8.1. Principal Component Analysis

To perform principal component analysis of the ancient individuals and the European populations from the Human Origins data set, we first conducted separate PCAs for the modern populations using all SNPs and each of the ancient individuals using the merged SNP data. Procrustes analysis [39,81] was then used to transform each individual's PC1 and PC2 loadings to the coordinate system of the PCA using all SNPs (setting the original coordinates of the ancient individual in that PCA to (0,0)). PC1 and PC2 loadings for the reference individuals were then averaged across all separate analyses. The individual PCAs for each ancient individual are shown in Figure S5.
A random allele was selected for each heterozygous modern individual, which made the data set completely homozygous. The smartpca tool (version 10210) of the EIGENSOFT package [82] was used for all PCAs, randomly excluding one SNP of a pair of SNPs with a linkage disequilibrium of r^2>0.7. Figure 1C shows kernel densities of the locations of modern populations with at least 10 individuals in the PC1-PC2 space. The 90% distribution range of these populations was estimated using the R package adehabitat [83] and plotted using GNU R [84].

We repeated this analysis including North African populations in order to look for any additional component of the Iberian farmers to modern North Africans (Figure S6). PC1 separates North Africans from Europeans while PC2 seems to be correlated with the amount of Near Eastern ancestry. All ancient samples line up along this gradient with one extreme in Druze and the other in Mesolithic Europeans. Farmers from El Portalon and Sweden are slightly shifted towards hunter-gatherers in comparison to central European farmers. The PCA suggests no additional North African ancestry in any of the ancient farmers.

An additional PCA including modern populations from the Caucasus was conducted since ADMIXTURE results (Supplementary Material S9) suggest some Eastern ancestry in some samples. PC1 correlates largely with Near Eastern ancestry with Druze and Mesolithic Europeans as the two extremes (Figure S7). PC2 has Sardinians and Tajiks as extremes suggesting some correlation with longitude. Ancient farmers group around Sardinians and the Chalcolithic El Portalon individuals form a line between Sardinians and Basques whereas central European farmers are shifted towards Near Eastern populations. There is no specific affinity to modern-day Caucasian populations for any of the ancient individuals.

S8.2. Outgroup f_3-statistics

We calculated an an outgroup f_3-statistic ([76,79], f_3(O; A, B)) in order to obtain a measurement of genetic relationship between two populations which is independent of an excess of drift in either of the populations. The statistic was calculated as:

$$f_3(\mathrm{O;A,B}) = \frac{\sum (p_O - p_A)(p_O - p_B) - (p_O - p_O^2)/(n_O - 1)}{\sum 2\, p_O (1 - p_O)}$$

where p_O is the allele frequency of the reference allele (arbitrarily chosen among the two alleles present at locus *i*) and n_O is the number of gene copies in population *O* (the outgroup) at locus *i*, with corresponding notations for populations *A* and *B*. In the absence of admixture with the outgroup (which we assume is the case), we expect the value of the statistic to be positive. A positive deviation from zero will be a function of the shared genetic history of two populations *A* and *B* in their unrooted population history with the outgroup *O.*

S8.3. D-statistics

To test deviations from a tree like population topology, we estimated the *D*-statistics [79]:

$$D(\mathrm{A,B;X,Y})=\frac{\sum_{\mathrm{i}=1}^{n}\left[(p_{\mathrm{iA}}-p_{\mathrm{iB}})(p_{\mathrm{iX}}-p_{\mathrm{iY}})\right]}{\sum_{\mathrm{i}=1}^{n}\left[(p_{\mathrm{iA}}+\mathrm{p}_{\mathrm{iB}}-2p_{\mathrm{iA}}p_{\mathrm{iB}})(p_{\mathrm{iX}}+\mathrm{p}_{\mathrm{iY}}-2p_{\mathrm{iX}}p_{\mathrm{iY}})\right]}$$

where p_{iA} is the frequency of one allele (arbitrarily chosen from the two alleles present) in population *A* at marker *i.* Numerator and denominator are summed across all n markers, we then obtained standard errors by performing a block jackknife over blocks of 500 SNPs in the genome. Significant deviations from 0 can be interpreted as deviations from a tree like population history of the form (*A*,*B*)(*X*,*Y*) due to gene flow, where positive values suggest excess affinity between A and X and/or Y and B and negative values excess affinity between A and Y and/or B and X. Using an outgroup population (e.g. chimpanzee or Sub-Saharan Africans) as population A narrows these options down to gene flow between B and X (in case of positive values) or between B and Y (if values are negative).

S8.4. Diversity estimates

To obtain an estimate of population diversity, we estimated a scaled diversity for all ancient populations with at least two contemporary individuals [20]. We require two different individuals since low coverage sequencing does not allow to call heterozygous sites confidently and we do not know whether particular individuals might be the result of recent inbreeding which would cause lower diversity estimates. We only used the transition-free merged sequence data set and not the genotype data in order to avoid ascertainment bias, post-mortem damage and to increase the number of sites. For each SNP site ascertained in the Yoruban population, we randomly sampled one read from each ancient individual. Our diversity estimate is the average number of mismatches across all sites with coverage in both individuals. We estimated standard errors for this statistic using a block jackknife procedure over blocks of 500 SNPs.

Diversity was estimated for all sites or cultures with two reasonably contemporary individuals and decent coverage: sites Ajvide (using Ajv58 and Ajv70), Motala (Motala12 and Motala1), Gökhem (Gok2 and Gok4), El Portalón (ATP2 and ATP12-1420) and the culture Alföld Linear

Pottery (ALP; NE1 and NE5). This procedure was chosen to avoid the effects of potential inbreeding. The Scandinavian hunter-gatherers show the lowest diversity of all groups whereas the Scandinavian farmers from Gökhem are intermediate between those and the central European and Iberian farmers (Figure S9). Generally, farmers show a higher diversity than hunter-gatherers which is consistent with previous results and might be attributed to the increased carrying capacity of farming groups and/or the admixture with hunter-gatherers [19,20].

S9. Admixture graph inference

We inferred a bifurcating population history that allows for directed pulse admixture events using TreeMix v1.12 [98]. TreeMix estimates a maximum-likelihood tree from the covariance matrix of allele frequencies and then adds migration edges to account for residual covariance. We selected the high coverage Denisovan genome (used as root), Yorubans from the 1000 genomes project (used for SNP ascertainment), MA1 (as representative of ancient North Eurasians) and the highest coverage individual from each of the ancient European groups (ATP2, Gok2, Ajv58, NE1, Loschbour, Stuttgart, Motala12, Iceman, LaBrana)[19,20,71,73,74]. Overlapping sequence data was available for these individuals at 396,797 transversion SNPs. Since all groups were represented by single individuals, correction for low samples sizes was turned off (-noss) and standard errors were estimated using blocks of 500 SNPs. All graphs in all settings were estimated with 100 different random seeds. The majority of runs (between 75% and 90% for all models) supported the results we discuss below.

The maximum-likelihood tree without migration edges clearly shows that early European farmers and late hunter-gathers form two distinct groups, with a further split of the hunter-gathers into a Western (Loschbour, LaBrana) and a Scandinavian clade (Motala12, Ajvide58) which is consistent with previous results [19,20]. The Upper Paleolithic Siberian MA1 forms an outgroup to these European groups.

The highest residual covariance of the maximum-likelihood exists between MA1 and Motala12 which is resolved by the first migration edge from Motala12 into MA1 (45.6%). Shared ancestry between these two groups was observed before, and it is unclear if this shared ancestry represents migration from the east followed by admixture [19] or clinal shared ancestry across Eurasia in the Paleolithic without direction. The second migration event is indicated from Ajv58 into Gok2 (32.5%), a signal that has been reported before [20], but we can confirm that the Scandinavian Neolithic hunter-gatherers are better representatives for the source of this migration into Scandinavian Neolithic farmers than Scandinavian Mesolithic hunter-gatherers. The third migration edge accounts for shared ancestry between common ancestor of Loschbour and LaBrana, and ATP2 (30.0%). The observation that the hunter-gatherer related ancestry into Gok2 and ATP2 originates from different sources is consistent with the results of D-tests (e.g. D(Mbuti, ATP2; Ajv58, Loschbour) > 0, |Z| > 3). The fourth migration event models admixture from Ajv58 into Iceman (18.7%) (Figure S10). Allowing for higher number of migrations did not produce any significant migration edges.

To summarize, we observe several migration edges from hunter-gatherer source populations into farming groups (Table S10). Notably, these sources differ for the farmer populations suggesting that several local admixture events happened. Chalcolithic and Scandinavian Neolithic farmers (ATP2, Gok2, Iceman, which are all dated to approx. 5000 BP) seem to harbor a higher proportion of hunter-gather related ancestry than the first Neolithic farmers of central and eastern Europe (Stuttgart, NE1).

We note that some of the remaining runs led to graphs with better likelihoods (no topology occurred in more than 4% of all runs). Two common migrations in these cases were into or from Yorubans and admixture from basal hunter-gathers (LaBrana, Loschbour) into more basal farming populations (ATP2, Gok2, Iceman) with a back-migration from Gok2 into Ajv58 to account for their residual shared ancestry among those two. Migrations involving Yorubans do not affect our general conclusions whereas the latter observation suggests a slightly different picture of a general admixture among both groups probably in central or Eastern Europe followed by local admixture in Scandinavia. An excess of shared ancestry between Gok2 and Ajv58 is consistent with the results of the D-tests. However, the D statistics also show that Gok2 and ATP2 have higher affinities to hunter-gatherers than Iceman which would not be consistent with these rare outlier TreeMix results.

The temporal differences between the Neolithic and Chalcolithic samples account for more than 1,000 years and they might have an influence on admixture estimates since more time for potential admixture had passed for the Chalcolithic individuals. Therefore, we repeated the TreeMix analysis only using the almost contemporary Gok2, Iceman, ATP2 and the Chalcolithic Hungarian CO1 as representatives for farmers, and Ajv58, LaBrana and Loschbour as representatives for hunter-gatherer populations. The Denisovan genome and Yorubans were included as outgroups. The merged data set contained 348,663 transversion SNPs without missing data in any of these samples. TreeMix was run as described above but migrations were inferred in a more supervised version as we pre-defined migration events from all hunter-gatherers into all Chalcolithic farmers (-cor_mig with starting value 0.0). The 12 migration events are all significant (Table S11, $p<0.05$) and the results are qualitatively similar to the results of the unsupervised analysis:

1. All four farmers receive significant hunter-gatherer ancestry from different source populations indicating that admixture happened continuously in different parts of Europe during and after the spread of farming.

2. Gok2 receives the highest total percentage of hunter-gatherer related ancestry (~32%), followed by ATP2 (~28%), Iceman (~23%) and CO1 (~20%). Which confirms that all these groups are substantially admixed and it seems like populations at the fringe of the Neolithic expansions mixed to a greater extent than farmer groups in central and eastern Europe.

3. As suggested by the unsupervised analysis, Gok2 and Iceman likely mixed with a population related to Ajv58, whereas ATP2 receives similar proportions of genetic material from LaBrana and Loschbour. CO1 receives similar proportions from Ajv58 and LaBrana, and a little less from Loschbour. One possible explanation for the latter observation might be that we do not have hunter-gatherer samples from Eastern Europe which would be better representatives for the populations that mixed with the Hungarian samples.

S10. Model-based clustering

Individuals were clustered using an unsupervised clustering algorithm as implemented in ADMIXTURE [85]. ADMIXTURE assigns proportions of each individual's genome to one of K (user defined number of clusters) ancestral populations. We included all Western Eurasian and North African populations from the Human Origins data set [19,79] and all Mesolithic, Neolithic and Chalcolithic individuals with more than 0.3x genome coverage.

All ancient individuals were merged with the Human Origins data (as described in Supplementary Material S7). The data set was then thinned for linkage disequilibrium using PLINK 1.07 [86]. Pairwise LD was calculated in windows of 200 SNPs with a step size of 25 SNPs and an r2 threshold of 0.4. We explored between 2 and 15 clusters using 50 replicates per K. Common signals between the different replicates for K between 2 and 15 were identified using the LargeKGreedy mode of CLUMPP [87]. Clustering was visualized using distruct [88].

Our results for modern populations are highly consistent with previous studies of Eurasian population structure [19,89–91]. For low numbers of clusters, we first see a distinction of Eastern Asia followed by additional clusters in Southern and Western Asia as well as North Africa (K=2 to K=7). K=8 introduces a new cluster which occurs in substantial frequency in the Caucasus and Kalash, but also in low levels for different European populations. This component is likely related to the ancient North Eurasian component identified by [19] which might have been introduced during the Bronze Age [68,92]. We chose to display K=10 in the main paper (Figure 3) since it is the lowest value of K to show a clear distinction between the Mesolithic European hunter-gatherer component (blue, modal in Northern Europeans), early farmers (orange, modal in Sardinians), North Africans (yellow) and distinct Near Eastern ancestry (white). Higher values of K work out more regional population structure in different parts of Eurasia.

S11. Affinity with other populations

In order to resolve the relationship of the four higher coverage El Portalón individuals (ATP2, ATP9, ATP12-1420, ATP9), we conducted D-tests [79,93] to test different tree topologies. We report tests with |Z|>2 as significant but we have to note that this threshold would not imply significance if we would correct for multiple testing and, thus, we likely report some false positives. Consistent patterns among related populations (which could be grouped to reduce multiple tests), however, suggest genuine signals in many cases.

S11.1. Relationship to other early European farmers

PCA (Figure 1C) suggests slightly greater genetic similarities between the Portalón individuals and Gok2. We also used D-tests to check whether the Iberian early farmers and the Scandinavian early farmer form a group to the exclusion of central European early farmers. The results are shown in Dataset S3. Only two out of 12 topologies of the shape (Mbuti, Central European early farmer; Portalón, Gok2) were rejected (|Z|>2) so most of them are consistent with the data. Both of the topologies that are rejected involve Iceman as outgroup (D(Mbuti, Iceman; ATP9, Gok2), Z=2.56 and D(Mbuti, Iceman; ATP2, Gok2), Z=3.30) and are rejected because of additional affinity between Iceman and Gok2. This result is likely caused by a similar source of the hunter-gatherer admixture in Gok2 and Iceman.

However, when we test topologies of the shape (Mbuti, Gok2; Portalón, Central European early farmer) or (Mbuti, Portalón; Gok2, Central European early farmer) we need to take into account that these tests will be highly affected by the different hunter-gatherer components, especially in Late Neolithic/Chalcolithic individuals that contain a greater fraction HG ancestry. We restrict such tests to later farmers (Chalcolithic and Scandinavian Neolithic farmers, all dated to about 5000 years BP). Three topologies out of eight with Gok2 as outgroup are rejected but only one of them (D(Mbuti, Gok2; ATP2, CO1), Z=-2.74) because of affinities between Gok2 and a Portalón individual. The remaining two topologies are rejected because of the affinities between Iceman and Gok2 (D(Mbuti, Gok2; ATP9, Iceman), Z=2.69 and D(Mbuti, Gok2; ATP9, Iceman), Z=2.07). Furthermore, only one out of eight topologies with a Portalón individual as an outgroup is not consistent with the data (D(Mbuti, ATP2; Gok2, Iceman), Z=-2.6). Consequently, most topologies including only later farmers are consistent with the data and D-tests do not give a clear picture whether Gok2 and Portalón individuals form a separate group from Chalcolithic central European farmers. This could be due to a lack of resolution and power (low coverage data from single individuals) or because the population structure of Neolithic Europe is more complicated than suggested by the PCA.

S11.2. Relationship to ancient European hunter-gatherer populations

Admixture from hunter-gatherers into incoming farmers has been shown in Scandinavia and Central Europe [19,20]. It remains an open question whether this admixture was a single event, most likely in South-Eastern Europe or whether mixing between the groups was ubiquitous in different parts of Europe. Our analysis using admixture graphs suggests different sources for the admixture into Iberian, Central European and Scandinavian farmers (Supplementary Material S9). To test different sources of hunter-gatherer admixture into early farmers, we conducted D-tests of the shape D(Mbuti, farmer; hunter-gatherer 1, hunter-gatherer 2). In this case we used – as others before [68,92] - KO1 (individual found in a farming context) as a proxy for Hungarian hunter-gatherers since he grouped with Mesolithic individuals in all other analyses. The results are shown in Dataset S4. The highest proportion of Mesolithic ancestry in the Portalón individuals seems to be related to central European hunter-gatherers (KO1, Loschbour) and not to the geographically close LaBrana (several |Z| >2). Central European farmers (CO1, Iceman, NE1) exclude only Mesolithic Scandinavians (Motala12) as a possible source so it seems likely that their admixture happened in Central Europe as well. Surprisingly, Mesolithic Scandinavians (Motala12) are excluded as a possible source of admixture into the Scandinavian farmer Gok2 whereas all other hunter-gatherer groups (including the Neolithic Scandinavian Ajv58) are consistent with the data. This suggests multiple admixture events into Scandinavian farmers which happened in different parts of Europe. However, we note that the currently available data does not allow us to detect a strong population structure in Mesolithic Europe. Only Scandinavia seems to be an outlier from a relatively uniform Mesolithic population.

Since the analysis above was based on only ~100,000 transversion SNPs per individual (from the Human Origins array), we might have limited power to detect significant differences. To increase that power, we utilized the sequence data used for the admixture graph analysis (Supplementary Material S9) and conducted a similar analysis. This allows us to use up to 1.9 million transversion SNPs per comparison. In this case, we tested D(Yorubans, farmer; hunter-gatherer 1, hunter-gatherer 2) with five different farmers and four hunter-gatherers (Table S12). The main results are similar although some signals are clearer. ATP2 and even the Scandinavian farmer Gok2 exclude Motala12 as potential source of admixture (|Z|>2.2). In the case of ATP2, all other hunter-gatherers (Loschbour (|Z|=4.5), LaBrana (|Z|=2.9) but also Ajv58 (|Z|=2.2) exclude Motala but not each other (|Z|<2)) are possible sources, whereas Gok2 seems to obtain most of its hunter-gatherer ancestry from a population related to Ajv58 (|Z|>2.4). The source of admixture into the Tyrolean Iceman is also more likely Ajv58 than Loschbour or Motala12 (|Z|>2.1) which is consistent with the results of Treemix (Supplementary Material S9). Furthermore, these results suggest that the Neolithic Scandinavian hunter-gatherer Ajv58 is closer to the different early farmers than the Mesolithic Scandinavian Motala12 (|Z|>2.2 for all but NE1 and Stuttgart), likely as a consequence of hunter-gatherer gene flow into Chalcolithic and Scandinavian Neolithic farmers. NE1 and Stuttgart are the oldest farmers in this analysis and they also consistently showed the lowest proportion of HG-related ancestry in all of our analyses. NE1 shows no exclusive affinities to any of the hunter-gatherers (|Z|<1.97) while Stuttgart only shows higher affinities to LaBrana than to Ajv58 (|Z|=2.21).

In order to measure the enrichment of HG-related ancestry in farmers over time, we calculated f4 statistics of the shape (Yoruban, ref_HG; ref_farmer, farmer) for the ancient individuals merged with sequence data, and correlated it with the age of the individuals (Figure 2B). We chose ref_HG to be one European Mesolithic individual with more than 1x sequencing depth and ref_farmer as one early Neolithic farmer from central Europe requiring more than 1x sequencing depth. We show the results using Loschbour and Stuttgart as references in the main text since they are the highest coverage individuals. Using different reference individuals or genotype data, however, had no substantial influence on the qualitative results, but it also suggests that Mesolithic Scandinavian hunter-gatherers are a poorer fit for the HG-related component in most farmers (Table S13).

S11.3. Affinity to modern-day North African populations

The geographic proximity of Iberia to Northern Africa opens up possibilities to migrations across the Strait of Gibraltar. In fact, farming reached Northern Africa and Southern Spain long before Northern Iberia, and modern Iberian populations show a significant proportion of North African ancestry [94]. Admixture estimates and outgroup f_3 statistics do not support a strong contribution of North African populations to the individuals of El Portalón. Modern-day North Africans are highly admixed with contributions from Europe, sub-Saharan Africa, the Near East [95] and Neandertals [96], and the level of admixture vary among groups. In order to avoid other components in reference populations from confounding the D-tests, we assume that all early European farmers contain the same Near Eastern component (which is also found in North Africa to some degree) and conduct D-test in the form of (Mbuti, modern-day North African; ancient farmer 1, ancient farmer 2). We use Mozabite, Saharawi, Algerian, Tunisian and Burbur as representatives of modern-day North African populations since the a particular ancestry component (the ‘North African component’) is maximized in these groups in the admixture analysis (Dataset S1). These analyses (Dataset S5) demonstrate that ATP2 and ATP12-1420 have similar genetic affinities to North Africans as Central European early farmers have ($|Z|<1.4$). However, ATP16 shows higher affinities to North Africa than other ancient farmers ($|Z|>2$ for NE1, CO1 and Iceman), suggesting that there was at least some contribution from North Africa ~5,000 years ago (in one out of eight Portalón individuals). Surprisingly, ATP9 shows the lowest North African affinity of all ancient farmers (all D(Mbuti, North African; ATP9, other farmer)>0, Z>2 for Iceman, Gok2, Stuttgart). Since ATP9 also represents the youngest individual (Bronze Age) in the analysis, we suspect that this is the result of increased admixture with other European groups in the Bronze age, which contained less North African or Near Eastern ancestry. Generally, genomic data from Neolithic North Africans is needed to solve the question whether there was a strong Neolithic African contribution to the Iberian Neolithic population.

S11.4. Relationship to modern-day Spanish populations

The PCA (Figure 1C) and outgroup f_3 statistics indicate that the El Portalón cluster close to modern Spanish populations but they also show general affinities with southwestern Europeans. The closest group on the Iberian Peninsula are Basques (both people living in the Basque country, Pais Vasco, and self-identifying Basques from France and Spain). This supports the observation from PCA and ADMIXTURE that Basques are closely related to the Chalcolithic farmers in the region. We conducted additional D-tests to check if Basques (Spanish Basques, French Basques and people from Pais Vasco) and El Portalón individuals form a group to the exclusion of other modern Spanish populations (Dataset S6). ATP2 groups exclusively with the Basque populations, all other ingroups are rejected (Z>2.5, corresponding to a FDR<0.01, [97]) and using all other Spanish populations as an outgroup for the grouping of ATP2 and Basques is consistent with the data. Only one (out of 36) topology of the shape (Mbuti, Basque; other Spanish, ATP12-1420) is not rejected (using Aragon instead of Pais Vasco as ingroup, Z=1.92). Aragon is an autonomous community in the North of Spain, likely with more shared ancestry with Basques compared to groups from other regions. Furthermore, (Mbuti, Galicia; ATP12-1420, Basque_French) (Z=-2.06) and (Mbuti, Galicia; ATP12-1420, Pais Vasco) (Z=-2.23) are rejected because of additional affinity between ATP12-1420 and Galicians. Galicians as an outgroup are still the better fit since the reversed topology is strongly rejected (|Z|>4). All 36 topologies with Basques as an outgroup to ATP16 are rejected. However, the small African contribution to ATP16 (see above) might be the cause of the rejection of several topologies (with Andalucia, Aragon, Baleares, Castilla La Mancha, Castillay Leon, Cataluna, Galicia, Murcia as outgroups to (Basques, ATP16), |2|<Z<|3|) since most modern Spanish populations carry more North African ancestry than Basques. Due to the lower coverage (0.41x compared to >1x for ATP2, ATP12-1420, and ATP16) and the high similarity among Spanish populations, we excluded ATP9 from this analysis since most topologies were consistent for ATP9 (probably due to a lack of power).

To conclude, most topologies grouping the El Portalón individuals with Basques (to the exclusion of other Iberian groups) are consistent with the data. There are, however, some exceptions which can be attributed to three reasons: (i) heterogeneity among the El Portalón individuals from different centuries, (ii) the complex admixture history of different Spanish populations and (iii) the fact that – despite all similarity – Basques probably do not represent absolute continuity since the Chalcolithic people who lived in El Portalón.

To gain more insights into the relationship of ancient farmers to modern-day Basques and to utilize the additional power obtained from sequence data, we extracted five unrelated individuals from the Basque Country (Pais Vasco) from the 1000 genomes data (phase 3 [28], downloaded from ftp.1000genomes.ebi.ac.uk). These individuals are the same as some of those genotyped for the Human Origins data. The data was merged with Yorubans (for SNP ascertainment) and five ancient farmers as described in Supplementary Material S7. We then calculated D statistics of topologies (Yorubans, Spanish_Pais_Vasco; ATP2, early European farmer) in order to test whether the Chalcolithic El Portalón farmers represent the closest early farmers to modern-day Basques. The results are shown in Dataset S1. The results confirm that people living in Basque Country today are genetically more similar to El Portalón farmers than to the early Neolithic farmers (NE1, Stuttgart, |Z|>2.3 for all but ATP9 vs. Stuttgart). Basques also show higher affinity to El Portalón farmers than to later farmers (CO1, Iceman), these signals are only significant for the highest coverage Portalón individual (ATP2, |Z| >2), but the results from eight out of nine comparisons involving the other Portalón individuals point in the same direction (Dataset S1). Consistent with the PCA results (Figure 1C), Gok2 shows similar affinities with Basques when contrasted to the El Portalon farmers. The sign of the D statistic even suggests that Basques might be slightly closer to Gok2 (Z < 1.6). It is however unlikely that Gok2 contributed directly and substantially to modern Basques based on geography, and this particular (non-significant) indication is more likely to be cased by similar ratio of HG- and farmer-related components in Gok2 as in modern-day Basques.

S11.5. Relationship to other modern-day South-Western Europeans

Our results from ADMIXTURE, PCA and outgroup f3 statistics suggest that the El Portalón farmers are also closely related to other modern populations South-Western Europe namely Sardinians and French_South. We extended the analysis of S11.4 to these populations by testing different tree topologies with Basques, Sardinians and French_South as in- and outgroup.

Most topologies involving Sardinians were rejected. They were not consistent as an outgroup with French_South, Basque_Spanish, Basque_French and Pais Vasco being the ingroup. They were also rejected as an ingroup when French_South, Basque_Spanish, Basque_French and Pais Vasco (only for ATP2, but not ATP16 or ATP12-1420) were the outgroup. This suggests that the complex admixture patterns among these Southern Europeans are not consistent with a tree-like history.

The pattern for French_South is different since most topologies are accepted. French_South are rejected as both outgroup and ingroup when Sardinians are the other modern-day populations. Only one other topology is rejected: (Mbuti, Pais Vasco; French_South, ATP2). This likely shows power limitations due to the small sample size for the French and ancient populations. All results are shown in Dataset S7. The results generally confirm the strong affinity of the El Portalón farmers with modern-day South-Western Europe.

S12. Neolithization and Language

The Neolithic cultures appear in the Iberian Peninsula around 7,500 cal BP as a result of dispersal of human groups along the Mediterranean coastal areas and (eventually) visible as the Cardial culture in the western part of the Mediterranean (e.g. [102,103]). According to [104](see also papers in [105]) and others (e.g [106,107]), the dispersal of the Neolithic communities was related to the spread of Indo-European languages to Europe. However, this model (the Anatolian Hypothesis) coexists with a number of competing models. In particular, other models have placed the origin of Indo-European languages (or Proto-Indo-European) in the East, North of the Caspian and Black Seas, and in a chronologically younger period (often termed the Steppe Hypothesis, [108,109]). The recently confirmed eastern migration of human groups (linked to the Yamnaya group) into Europe around 4500 cal BP [68] has been interpreted as evidence for the Steppe Hypothesis, but the population movement is also consistent with a secondary expansion under the Anatolian Hypothesis. The linguistic implications of this are still under debate [106,110–112](e.g. [106,110–112]).

There are a number of different linguistic scenarios consistent with these genetic results, although as always it is difficult to associate genetic information confidently with archaeological groups or language families (see e.g. [68,113,114]). The Basque language (Euskara) is a linguistic isolate [115], and is believed to be the last surviving pre-Indo-European language in Western Europe ([116], see papers in [114]). The only known precursor to Basque is Aquitanian, reported during the Roman empire and spoken in southwest Gaul, the Pyrenees, and some adjoining areas [114,116]. This language is clearly related to Basque, but is probably a relative rather than the direct ancestor [115]. Basque has always been a magnet for extravagant linguistic speculation, but the hypothesis of Paleolithic roots of Basque is most wide accepted on the groups on the grounds of explanatory parsimony, and in the absence of adequate evidence for other hypotheses [115,116].

One intriguing suggestion is that the Basque language exhibits similarities to the pre-Roman language of Sardinia (Paleosardo) [117–119] based on, for example, place-names on Sardinia. The number of linguistic forms is small, but this is particularly interesting given Sardinians and Basques are the two modern populations with the highest genetic proportion of early farmer ancestry [19,20,74,77,120]. Contacts between Iberia and Sardinia in the Neolithic are indicated by recent studies of Obsidian artifacts [121], facilitated by maritime (and coastal) movement [102,103]. This suggests the Basque might be the remnant of a much larger Vasconic speaking area, suggesting a the possibility that language family spread along with the first farmers. If so it would be tempting to suppose that it was the *only* language of the first farmers, which would support the Steppe Hypothesis of Indo-European origins over the Anatolian Hypothesis.

Language isolates are however not uncommon outside of Europe. Of the approximately 350 language families in the world, 121 of are isolates [115]. The existence of such isolates is not really surprising given the highly skewed of linguistic diversity, and isolates are sporadically encountered embedded within the ranges of most large language families worldwide (e.g. [122]). Within the Indo-European languages, the Greek, Armenian and Albanian subgroups are also (near) isolates, consistent with the prediction of the Anatolian Hypothesis that the center of linguistic diversity in Europe would coincide with the entry points of the first farmers (and not contradicted by [68]). It is not implausible that Basque is an indigenous language that expanded in place after adoption of agriculture, or that Basque entered Europe alongside these other Indo-European languages. There is some hope that advances in Proto-Basque reconstruction will shed light onto these issues. Proposals of linguistic similarities between the Basque and other languages must however be evaluated with caution [114].

References

1. Bermúdez de Castro JM, Carbonell E, Arsuaga JL, editors. Gran Dolina Site: TD6 Aurora Stratum (Burgos, Spain). Journal of Human Evolution. 1999;37: 309–700.

2. Arsuaga JL, Martinez I, Arnold LJ, Aranburu A, Gracia-Tellez A, Sharp WD, et al. Neandertal roots: Cranial and chronological evidence from Sima de los Huesos. Science. 2014;344: 1358–1363. doi:10.1126/science.1253958

3. Carretero JM, Ortega AI, Juez L, Pérez-González A, Arsuaga JL, Pérez-Martínez R, et al. A Late Pleistocene-Early Holocene archaeological sequence of Portalón de Cueva Mayor (Sierra de Atapuerca, Burgos, Spain). Munibe (Antropologia-Arkeologia). 2008;59: 67–80.

4. VERGÈS J, ALLUÉ E, ANGELUCCI D, BURJACHS F, CARRANCHO A, CEBRIÀ A, et al. i VAQUERO, M.(2008):“Los niveles neolíticos de la cueva del Mirador (Sierra de Atapuerca, Burgos): nuevos datos sobre la implantación y el desarrollo de la economía agropecuaria en la submeseta norte.” A MS Hernández, JA Soler i JA López Padilla (eds): IV Congreso del Neolítico de la Península Ibérica (Alicante, 2006). pp. 418–127.

5. Pérez-Romero A, Carretero JM, Alday A, Arsuaga JL. La Cerámica Protohistórica e Histórica en el yacimiento del Portalón de Cueva Mayor, Sierra de Atapuerca. Burgos. Boletín de la Sociedad Española de Cerámica y Vidrio. 2013;52: 183–193.

6. Anderung C, Bouwman A, Persson P, Carretero JM, Ortega AI, Elburg R, et al. Prehistoric contacts over the Straits of Gibraltar indicated by genetic analysis of Iberian Bronze Age cattle. Proceedings of the National Academy of Sciences of the United States of America. 2005;102: 8431–8435.

7. Lira J, Linderholm A, Olaria C, Brandström Durling M, Gilbert MTP, Ellegren H, et al. Ancient DNA reveals traces of Iberian Neolithic and Bronze Age lineages in modern Iberian horses. Molecular ecology. 2010;19: 64–78.

8. Martínez-Pillado V, Aranburu A, Arsuaga JL, Ruiz-Zapata B, Gil-García MJ, Stoll H, et al. Upper Pleistocene and Holocene palaeoenvironmental records in Cueva Mayor karst (Atapuerca, Spain) from different proxies: speleothem crystal fabrics, palynology and archaeology. International Journal of Speleology. 2014;43: 1.

9. Delibes G, Herrán JI, Santiago J, Val J. Regional Views of Social Complexity. In: Lillios KT, editor. The origins of complex societies in late prehistoric Iberia. International Monographs in Prehistory; 1995.

10. Castilla M, Carretero J-M, Gracia A, Arsuaga J-L. Evidence of rickets and/or scurvy in a complete Chalcolithic child skeleton from the El Portalon site (Sierra de Atapuerca, Spain). Journal of anthropological sciences= Rivista di antropologia: JASS/Istituto italiano di antropologia. 2014;92: 257–271.

11. Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Bronk Ramsey C, et al. IntCal13 and Marine13 radiocarbon age calibration curves 0-50,000 years cal BP. 2013;

12. Rohland N, Hofreiter M. Ancient DNA extraction from bones and teeth. Nat Protoc. 2007;2: 1756–1762. doi:10.1038/nprot.2007.247

13. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci USA. 2013;110: 15758–15763. doi:10.1073/pnas.1314445110

14. Yang DY, Eng B, Waye JS, Dudar JC, Saunders SR. Technical note: improved DNA extraction from ancient bones using silica-based spin columns. Am J Phys Anthropol. 1998;105: 539–543. doi:10.1002/(SICI)1096-8644(199804)105:4<539::AID-AJPA10>3.0.CO;2-1

15. Svensson EM, Anderung C, Baubliene J, Persson P, Malmström H, Smith C, et al. Tracing genetic change over time using nuclear SNPs in ancient and modern cattle. Anim Genet. 2007;38: 378–383. doi:10.1111/j.1365-2052.2007.01620.x

16. Meyer M, Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harbor Protocols. 2010;2010: pdb.prot5448–pdb.prot5448. doi:10.1101/pdb.prot5448

17. Kircher M. Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol. 2012;840: 197–228. doi:10.1007/978-1-61779-516-9_23

18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

19. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513: 409–413. doi:10.1038/nature13673

20. Skoglund P, Malmstrom H, Omrak A, Raghavan M, Valdiosera C, Gunther T, et al. Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. Science. 2014;344: 747–750. doi:10.1126/science.1253448

21. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. Lalueza-Fox C, editor. PLoS ONE. 2012;7: e34131. doi:10.1371/journal.pone.0034131

22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

23. Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C. HAPLOFIND: A New Method for High-Throughput mtDNA Haplogroup Assignment. Human Mutation. 2013;34: 1189–1194. doi:10.1002/humu.22356

24. Van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30: E386–394. doi:10.1002/humu.20921

25. Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, et al. A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. Cell. 2008;134: 416–426. doi:10.1016/j.cell.2008.06.021

26. Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, et al. A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. Current Biology. 2013;23: 553–559. doi:10.1016/j.cub.2013.02.044

27. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, et al. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science. 2011;334: 94–98. doi:10.1126/science.1211177

28. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491: 56–65. doi:10.1038/nature11632

29. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

30. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 2015;43: D670–681. doi:10.1093/nar/gku1177

31. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

32. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15: 356. doi:10.1186/s12859-014-0356-4

33. Skoglund P, Storå J, Götherström A, Jakobsson M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. Journal of Archaeological Science. 2013;40: 4477–4482. doi:10.1016/j.jas.2013.07.004

34. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, et al. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet. 2012;90: 675–684. doi:10.1016/j.ajhg.2012.03.002

35. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 1999;23: 147. doi:10.1038/13779

36. Chandler H, Sykes B, Zilhão J. Using ancient DNA to examine genetic continuity at the Mesolithic-Neolithic transition in Portugal. Actas del III Congreso del Neolítico en la Península Ibérica: Santander, 5 a 8 de octubre de 2003. Servicio de Publicaciones; 2005. pp. 781–786.

37. Malmström H, Gilbert MTP, Thomas MG, Brandström M, Storå J, Molnar P, et al. Ancient DNA Reveals Lack of Continuity between Neolithic Hunter-Gatherers and Contemporary Scandinavians. Current Biology. 2009;19: 1758–1762. doi:10.1016/j.cub.2009.09.017

38. Malmstrom H, Linderholm A, Skoglund P, Stora J, Sjodin P, Gilbert MTP, et al. Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process. Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;370: 20130373–20130373. doi:10.1098/rstb.2013.0373

39. Skoglund P, Malmstrom H, Raghavan M, Stora J, Hall P, Willerslev E, et al. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science. 2012;336: 466–469. doi:10.1126/science.1216304

40. Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, et al. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science. 2009;326: 137–140. doi:10.1126/science.1176869

41. Der Sarkissian C, Balanovsky O, Brandt G, Khartanovich V, Buzhilova A, Koshel S, et al. Ancient DNA Reveals Prehistoric Gene-Flow from Siberia in the Complex Human Population History of North East Europe. Williams SM, editor. PLoS Genetics. 2013;9: e1003296. doi:10.1371/journal.pgen.1003296

42. Hervella M, Izagirre N, Alonso S, Fregel R, Alonso A, Cabrera VM, et al. Ancient DNA from hunter-gatherer and farmer groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe. PLoS ONE. 2012;7: e34417. doi:10.1371/journal.pone.0034417

43. Sánchez-Quinto F, Schroeder H, Ramirez O, Ávila-Arcos MC, Pybus M, Olalde I, et al. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. Current Biology. 2012;22: 1494–1499. doi:10.1016/j.cub.2012.06.005

44. Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, et al. Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoS Biol. 2010;8: e1000536. doi:10.1371/journal.pbio.1000536

45. Lacan M, Keyser C, Ricaut F-X, Brucato N, Tarrús J, Bosch A, et al. Ancient DNA suggests the leading role played by men in the Neolithic dissemination. Proc Natl Acad Sci USA. 2011;108: 18255–18259. doi:10.1073/pnas.1113061108

46. Lacan M, Keyser C, Ricaut F-X, Brucato N, Duranthon F, Guilaine J, et al. Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci USA. 2011;108: 9788–9791. doi:10.1073/pnas.1100723108

47. Gamba C, Fernández E, Tirado M, Deguilloux MF, Pemonge MH, Utrilla P, et al. Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. Mol Ecol. 2012;21: 45–56. doi:10.1111/j.1365-294X.2011.05361.x

48. Brandt G, Haak W, Adler CJ, Roth C, Szecsenyi-Nagy A, Karimnia S, et al. Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity. Science. 2013;342: 257–261. doi:10.1126/science.1241844

49. Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proceedings of the National Academy of Sciences. 2014;111: 4832–4837. doi:10.1073/pnas.1316513111

50. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, Jane Adler C, et al. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications. 2013;4: 1764. doi:10.1038/ncomms2656

51. Lacan M, Keyser C, Crubézy E, Ludes B. Ancestry of modern Europeans: contributions of ancient DNA. Cellular and Molecular Life Sciences. 2013;70: 2473–2487. doi:10.1007/s00018-012-1180-5

52. Gómez-Sánchez D, Olalde I, Pierini F, Matas-Lalueza L, Gigli E, Lari M, et al. Mitochondrial DNA from El Mirador Cave (Atapuerca, Spain) Reveals the Heterogeneity of Chalcolithic Populations. Hofreiter M, editor. PLoS ONE. 2014;9: e105105. doi:10.1371/journal.pone.0105105

53. Izagirre N, de la Rúa C. An mtDNA analysis in ancient Basque populations: implications for haplogroup V as a marker for a major paleolithic expansion from southwestern europe. Am J Hum Genet. 1999;65: 199–207. doi:10.1086/302442

54. Alzualde A, Izagirre N, Alonso S, Alonso A, De La Rúa C. Temporal mitochondrial DNA variation in the Basque Country: Influence of Post-Neolithic events. Annals of human genetics. 2005;69: 665–679.

55. Cardoso S, Alfonso-Sánchez MA, Valverde L, Odriozola A, Pérez-Miranda AM, Peña JA, et al. The maternal legacy of Basques in northern navarre: New insights into the mitochondrial DNA diversity of the Franco-Cantabrian area. American journal of physical anthropology. 2011;145: 480–488.

56. Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, Fornarino S, et al. Saami and Berbers—an unexpected mitochondrial DNA link. The American Journal of Human Genetics. 2005;76: 883–886.

57. Behar DM, Harmant C, Manry J, van Oven M, Haak W, Martinez-Cruz B, et al. The Basque Paradigm: Genetic Evidence of a Maternal Continuity in the Franco-Cantabrian Region since Pre-Neolithic Times. The American Journal of Human Genetics. 2012;90: 486–493. doi:10.1016/j.ajhg.2012.01.002

58. Cardoso S, Valverde L, Alfonso-Sánchez MA, Palencia-Madrid L, Elcoroaristizabal X, Algorta J, et al. The expanded mtDNA phylogeny of the Franco-Cantabrian region upholds the pre-neolithic genetic substrate of Basques. PloS one. 2013;8: e67835.

59. Gómez-Carballa A, Olivieri A, Behar DM, Achilli A, Torroni A, Salas A. Genetic Continuity in the Franco-Cantabrian Region: New Clues from Autochthonous

Mitogenomes. Caramelli D, editor. PLoS ONE. 2012;7: e32851. doi:10.1371/journal.pone.0032851

60. Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, Cuscó I, et al. New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS One. 2009;4: e5112.

61. Alfonso-Sánchez MA, Cardoso S, Martínez-Bouzas C, Peña JA, Herrera RJ, Castro A, et al. Mitochondrial DNA haplogroup diversity in Basques: A reassessment based on HVI and HVII polymorphisms. American Journal of Human Biology. 2008;20: 154–164. doi:10.1002/ajhb.20706

62. Van Geystelen A, Decorte R, Larmuseau MH. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. BMC Genomics. 2013;14: 101. doi:10.1186/1471-2164-14-101

63. Van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. Seeing the Wood for the Trees: A Minimal Reference Phylogeny for the Human Y Chromosome. Human Mutation. 2014;35: 187–191. doi:10.1002/humu.22468

64. Poznik GD, Henn BM, Yee M-C, Sliwerska E, Euskirchen GM, Lin AA, et al. Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females. Science. 2013;341: 562–565. doi:10.1126/science.1237619

65. Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, et al. Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs European Y-Chromosome Phylogeny. Science. 2013;341: 565–569. doi:10.1126/science.1237947

66. Regueiro M, Cadenas AM, Gayden T, Underhill PA, Herrera RJ. Iran: Tricontinental Nexus for Y-Chromosome Driven Migration. Human Heredity. 2006;61: 132–143. doi:10.1159/000093774

67. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Research. 2008;18: 830–838. doi:10.1101/gr.7172008

68. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015; doi:10.1038/nature14317

69. Rootsi S, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, et al. Phylogeography of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe. The American Journal of Human Genetics. 2004;75: 128–137. doi:10.1086/422196

70. Szecsenyi-Nagy A, Brandt G, Keerl V, Jakucs J, Haak W, Moller-Rieker S, et al. Tracing the genetic origin of Europe’s first farmers reveals insights into their social organization [Internet]. 2014 Sep. Report No.: 008664. Available: http://biorxiv.org/lookup/doi/10.1101/008664

71. Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, et al. Genome flux and stasis in a five millennium transect of European prehistory. Nature Communications. 2014;5: 5257. doi:10.1038/ncomms6257

72. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, et al. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014;346: 1113–1118. doi:10.1126/science.aaa0114

73. Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature. 2014;507: 225–228. doi:10.1038/nature12960

74. Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nature Communications. 2012;3: 698. doi:10.1038/ncomms1701

75. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514: 445–449. doi:10.1038/nature13810

76. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2013;505: 87–91. doi:10.1038/nature12736

77. Haak W, Brandt G, Jong HN d., Meyer C, Ganslmeier R, Heyd V, et al. Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age. Proceedings of the National Academy of Sciences. 2008;105: 18226–18231. doi:10.1073/pnas.0807592105

78. Lee EJ, Makarewicz C, Renneberg R, Harder M, Krause-Kyora B, Müller S, et al. Emerging genetic patterns of the european neolithic: Perspectives from a late neolithic bell beaker burial site in Germany. American Journal of Physical Anthropology. 2012;148: 571–579. doi:10.1002/ajpa.22074

79. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient Admixture in Human History. Genetics. 2012;192: 1065–1093. doi:10.1534/genetics.112.145037

80. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338: 222–226. doi:10.1126/science.1224344

81. Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, et al. Comparing spatial maps of human population-genetic variation using Procrustes analysis. Statistical applications in genetics and molecular biology. 2010;9.

82. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genetics. 2006;2: e190. doi:10.1371/journal.pgen.0020190

83. Calenge C. The package "adehabitat" for the R software: a tool for the analysis of space and habitat use by animals. Ecological modelling. 2006;197: 516–519.

84. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0; 2012.

85. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

86. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795

87. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007;23: 1801–1806. doi:10.1093/bioinformatics/btm233

88. Rosenberg NA. distruct: a program for the graphical display of population structure: PROGRAM NOTE. Molecular Ecology Notes. 2003;4: 137–138. doi:10.1046/j.1471-8286.2003.00566.x

89. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H-C, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008;451: 998–1003. doi:10.1038/nature06742

90. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science. 2008;319: 1100–1104. doi:10.1126/science.1153717

91. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456: 98–101. doi:10.1038/nature07331

92. Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522: 167–172. doi:10.1038/nature14507

93. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461: 489–494. doi:10.1038/nature08365

94. Botigue LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, et al. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proceedings of the National Academy of Sciences. 2013;110: 11791–11796. doi:10.1073/pnas.1306223110

95. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. Schierup MH, editor. PLoS Genetics. 2012;8: e1002397. doi:10.1371/journal.pgen.1002397

96. Sánchez-Quinto F, Botigué LR, Civit S, Arenas C, Ávila-Arcos MC, Bustamante CD, et al. North African Populations Carry the Signature of Admixture with Neandertals. Caramelli D, editor. PLoS ONE. 2012;7: e47765. doi:10.1371/journal.pone.0047765

97. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B. 1995;57: 289–300.

98. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8: e1002967. doi:10.1371/journal.pgen.1002967

99. Sverrisdottir OO, Timpson A, Toombs J, Lecoeur C, Froguel P, Carretero JM, et al. Direct Estimates of Natural Selection in Iberia Indicate Calcium Absorption Was Not the Only Driver of Lactase Persistence in Europe. Molecular Biology and Evolution. 2014;31: 975–983. doi:10.1093/molbev/msu049

100. Itan Y, Jones BL, Ingram CJ, Swallow DM, Thomas MG. A worldwide correlation of lactase persistence phenotype and genotypes. BMC Evolutionary Biology. 2010;10: 36. doi:10.1186/1471-2148-10-36

101. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, et al. Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet. 2010;6: e1000993. doi:10.1371/journal.pgen.1000993

102. Zilhão J. Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe. Proceedings of the National Academy of Sciences. 2001;98: 14180–14185.

103. Rowley-Conwy P. Westward Ho! Current Anthropology. 2011;52: S431–S451.

104. Renfrew C. Archaeology and language: the puzzle of Indo-European origins. CUP Archive; 1990.

105. Renfrew C, Bellwood P. Examining the farming/language dispersal hypothesis. Cambridge: McDonald Institute for Archaeological Research; 2002.

106. Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, et al. Mapping the origins and expansion of the Indo-European language family. Science. 2012;337: 957–960.

107. Gray RD, Atkinson QD. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature. 2003;426: 435–439.

108. Gimbutas M. Old Europe c. 7000-3500 BC: The earliest European civilization before the infiltration of the Indo-European peoples. Journal of Indo-European Studies. 1973;

109. Gimbutas M. The Kurgan Culture and the Indo-Europeanization of Europe: selected articles from 1952 to 1993. Study of Man; 1997.

110. Chang W, Cathcart C, Hall D, Garrett A. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. Language. 2015;91: 194–244.

111. Anthony DW, Ringe D. The Indo-European Homeland from Linguistic and Archaeological Perspectives. Annual Review of Linguistics. 2015;1: 199–219. doi:10.1146/annurev-linguist-030514-124812

112. Mallory JP. Twenty-first century clouds over Indo-European homelands. Journal of Language Relationship. 2013;9: 145–154.

113. Cavalli-Sforza LL. The Basque population and ancient migrations in Europe. Munibe. 1988;6: 129–137.

114. Hualde JI, Lakarra JA, Trask RL. Towards a history of the Basque language. John Benjamins Publishing; 1996.

115. Campbell L. Language Isolates and Their History, or, What's Weird, Anyway? Paper delivered at the 36th annual meeting of the Berkeley Linguistics Society, Berkeley, CA. 2010.

116. Trask RL. The history of Basque. Routledge; 2013.

117. Blasco-Ferrer E. Paleosardo: le radici linguistiche della Sardegna neolitica. Walter de Gruyter; 2010.

118. Blasco Ferrer EB. Iberian ordumeles, Paleo-Sardinian Ortumele, Ortarani and Araunele. Cognitive Semantics and Substrata Research. Journal of Indo-European studies. 2010;38: 373–383.

119. Morvan M. Les origines linguistiques du basque. Presses Univ de Bordeaux; 1996.

120. Sikora M, Carpenter ML, Moreno-Estrada A, Henn BM, Underhill PA, Sánchez-Quinto F, et al. Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. PLoS Genet. 2014;10: e1004353. doi:10.1371/journal.pgen.1004353

121. Terradas X, Gratuze B, Bosch J, Enrich R, Esteve X, Oms FX, et al. Neolithic diffusion of obsidian in the western Mediterranean: new data from Iberia. Journal of Archaeological Science. 2014;41: 69–78.

122. Lindström E, Terrill A, Reesink G, Dunn M. The Languages of Island Melanesia. Genes, Language, and Culture History in the Southwest Pacific: A Synthesis Oxford University Press, Oxford. 2007; 118–140.

123. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. Nat Genet. 2002;30: 233–237. doi:10.1038/ng826

124. Lamason RL, Mohideen M-APK, Mest JR, Wong AC, Norton HL, Aros MC, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science. 2005;310: 1782–1786. doi:10.1126/science.1116238

125. Graf J, Hodgson R, van Daal A. Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. Hum Mutat. 2005;25: 278–284. doi:10.1002/humu.20143

126. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007;39: 1443–1452. doi:10.1038/ng.2007.13

127. Zhang M, Song F, Liang L, Nan H, Zhang J, Liu H, et al. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. Hum Mol Genet. 2013;22: 2948–2959. doi:10.1093/hmg/ddt142

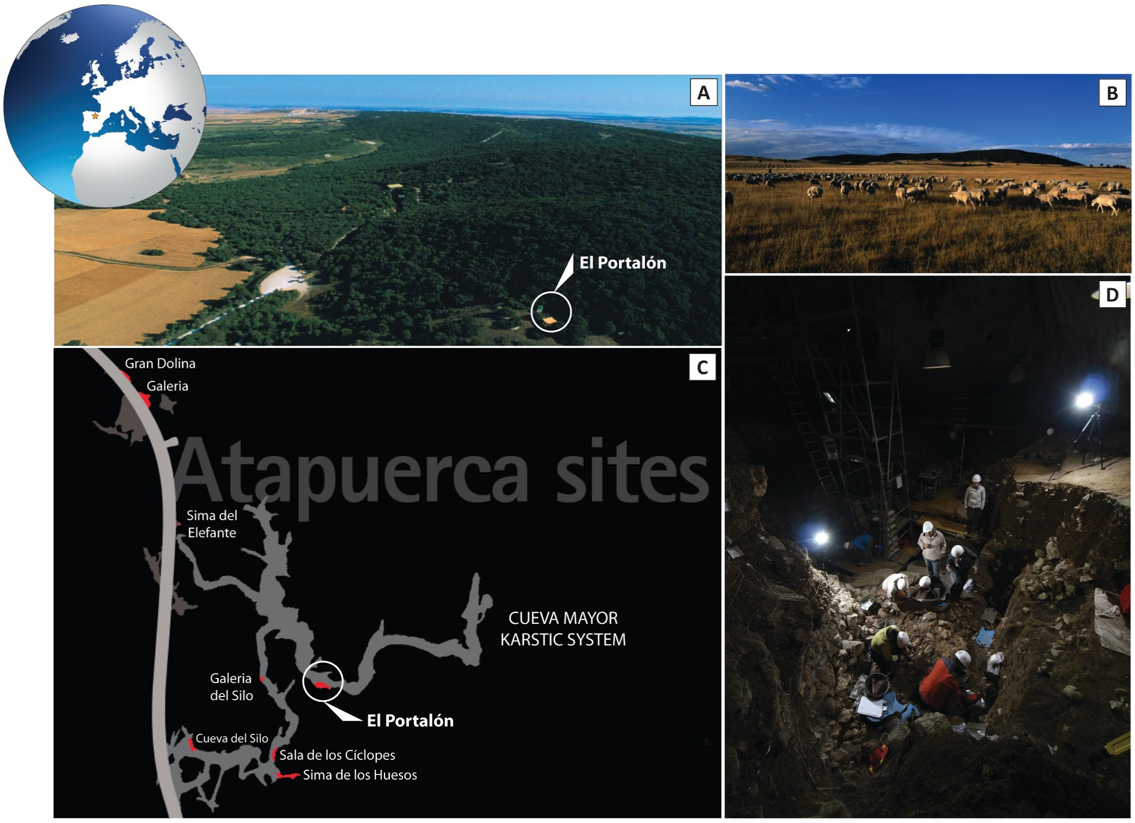


Figure S1: Location of the El Portalón archaeological site. (A and B) Panoramic views of the Sierra de Atapuerca with indication of the major archaeological sites including El Portalón. (C) Simplified map of the Cueva Mayor karstic system and the location of the El Portalón site. (D) View of the El Portalón site excavation area.

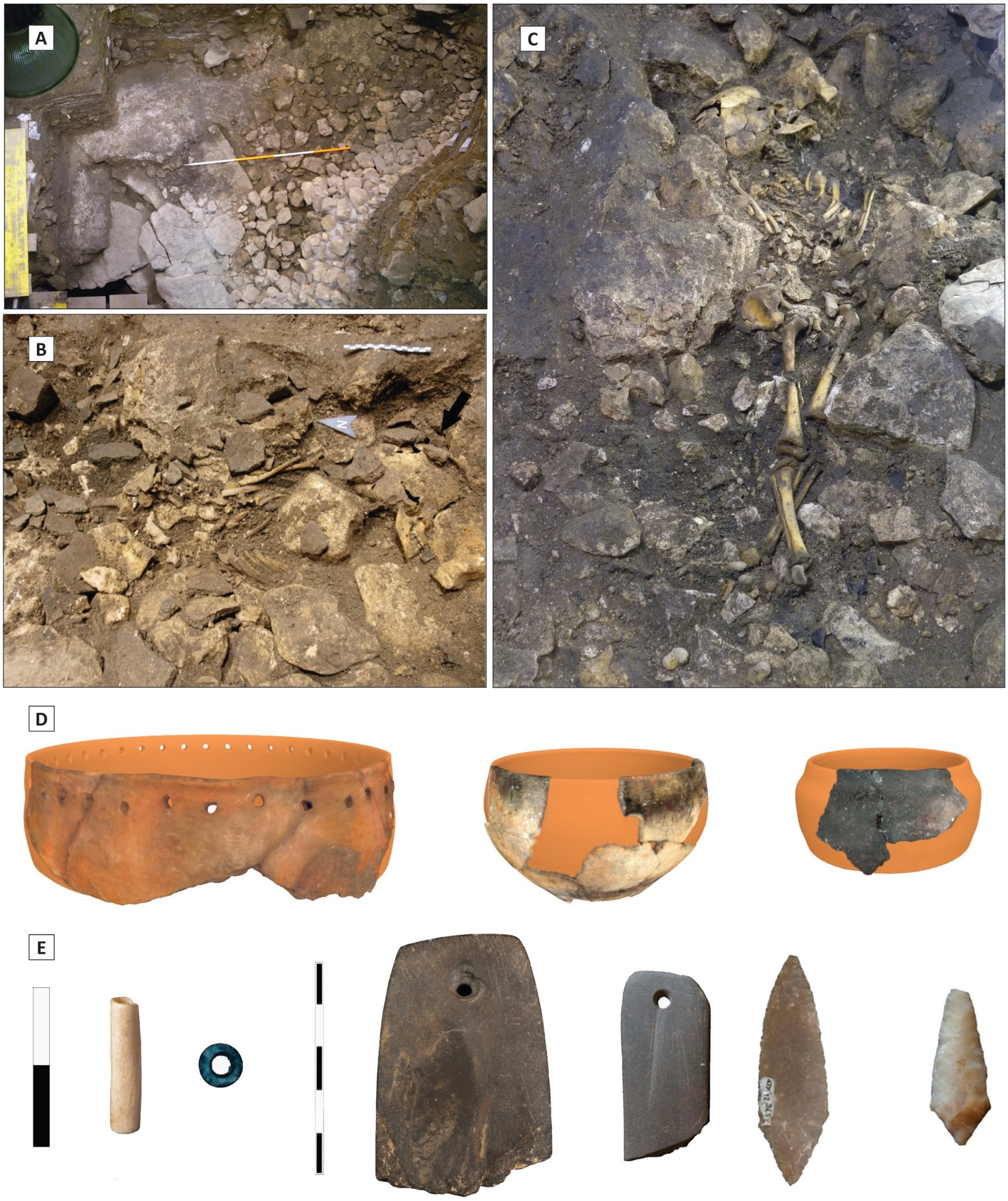


Figure S2: Chalcolithic funerary context in the El Portalón site. (A) View of the excavation area during the excavation of the tumular structure. (B) Detail of the ATP12-1420 burial with associated pottery fragments. (C) ATP-1420 burial. (D) Reconstruction of different pottery vessels associated to the burials. (E) Different grave good objects found in the burials: cylindrical and circular bone beads, two fragmented slate archery armguards and flint arrowheads.

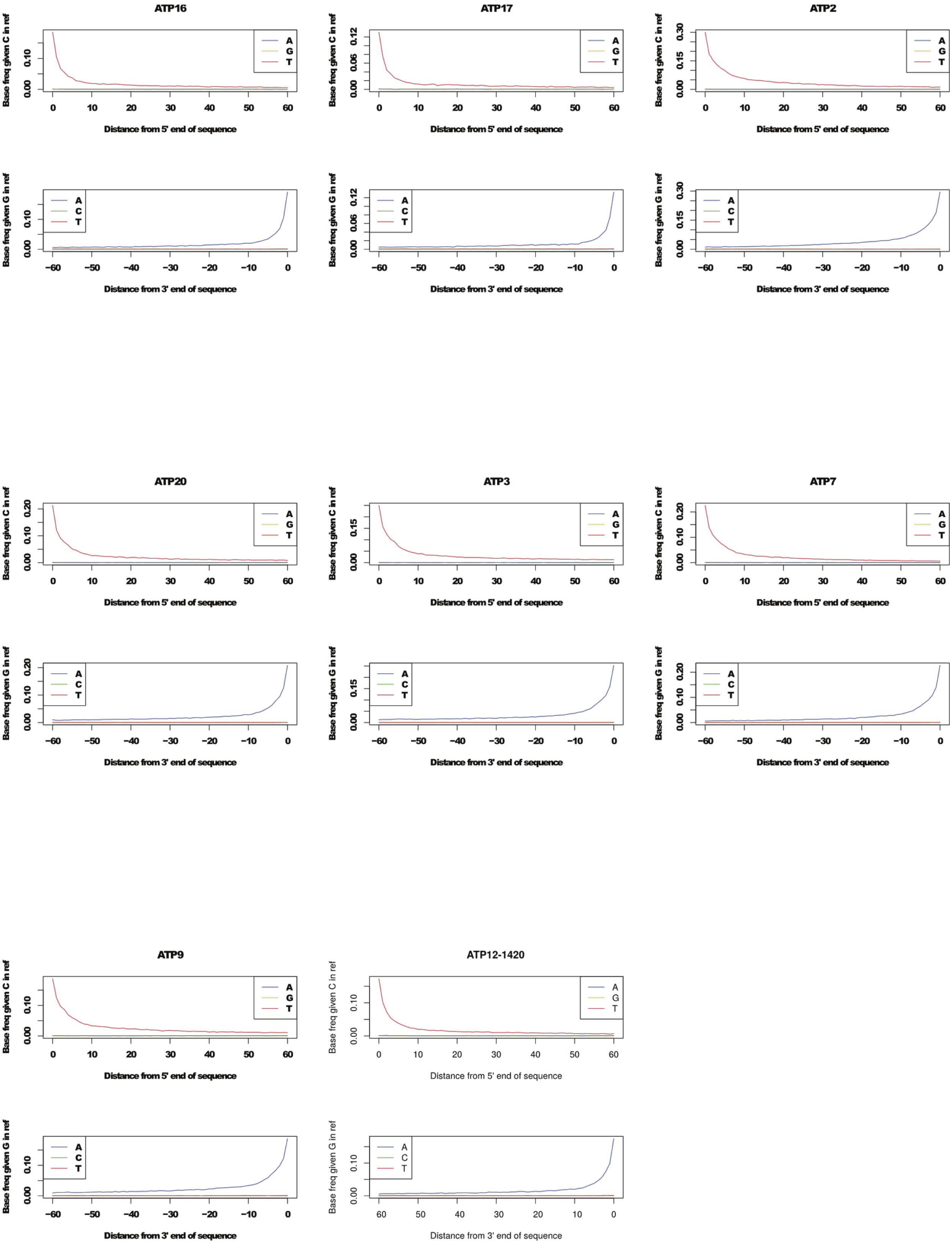


Figure S3: Deamination pattern for the eight samples sequenced for this study.

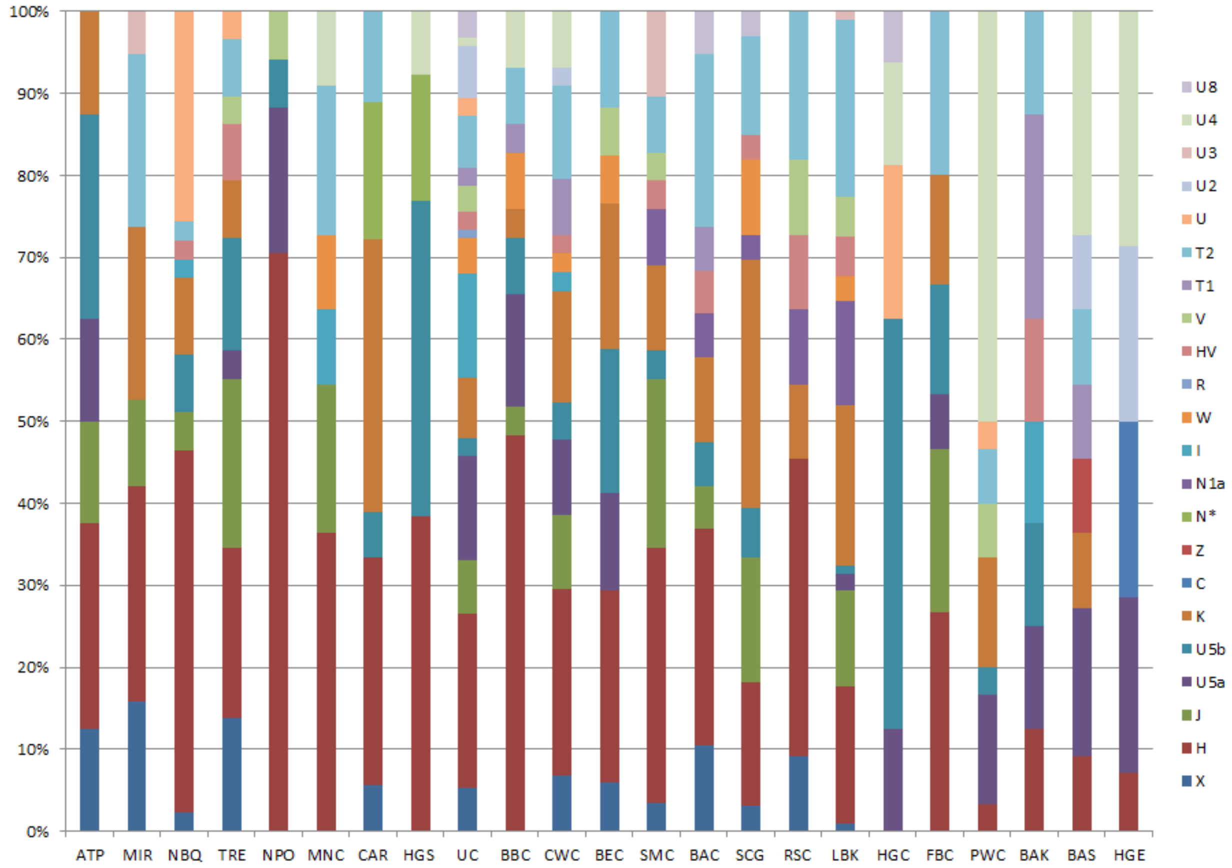


Figure S4: The prehistoric population haplogroup frequencies from Dataset S2 displayed as a histogram.

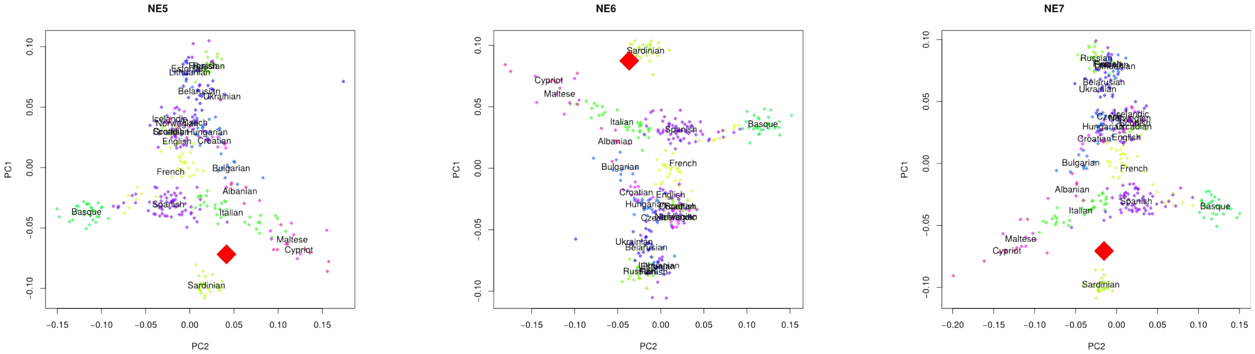


Figure S5: Individual PCAs for all ancient individuals together with present-day European populations.

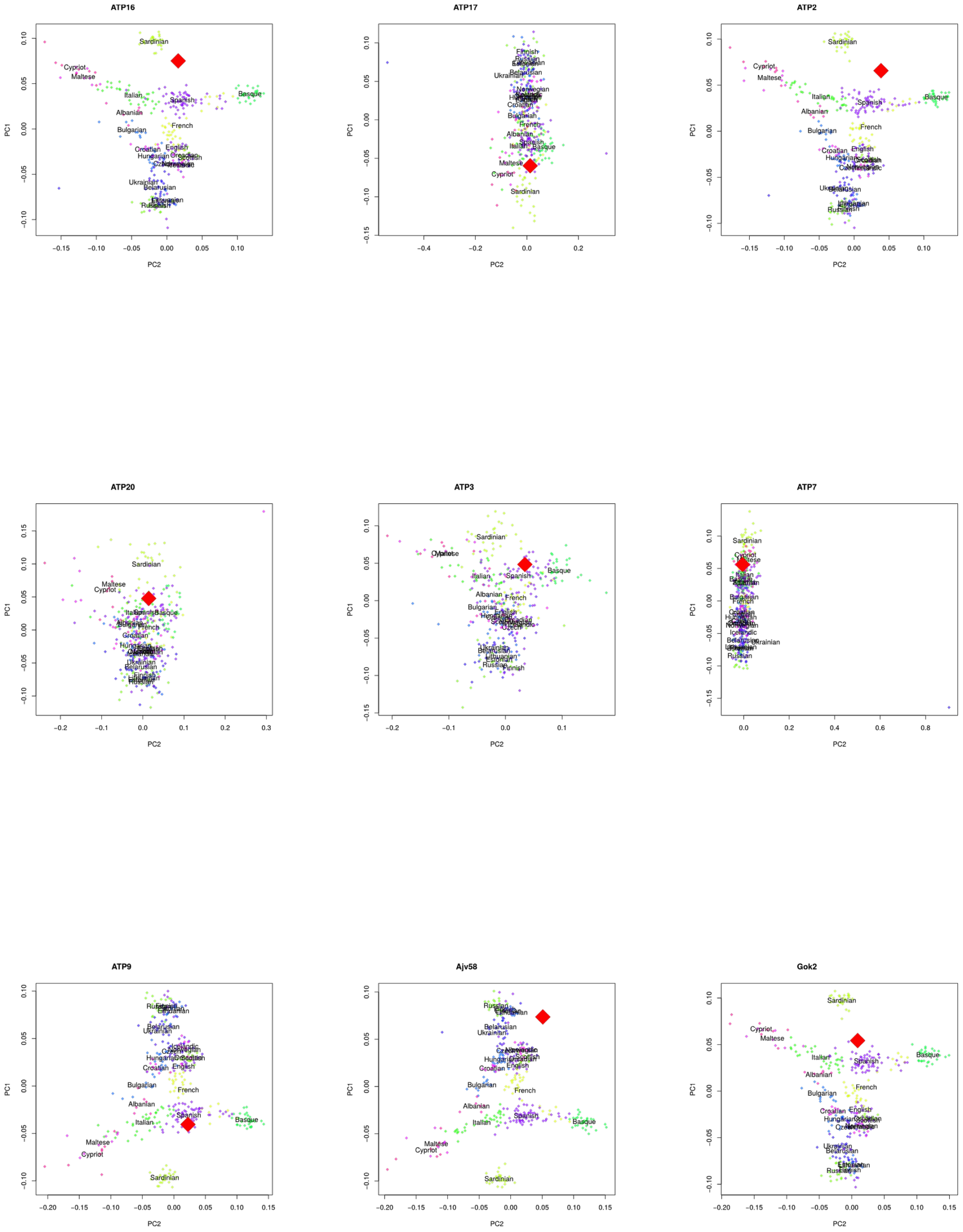
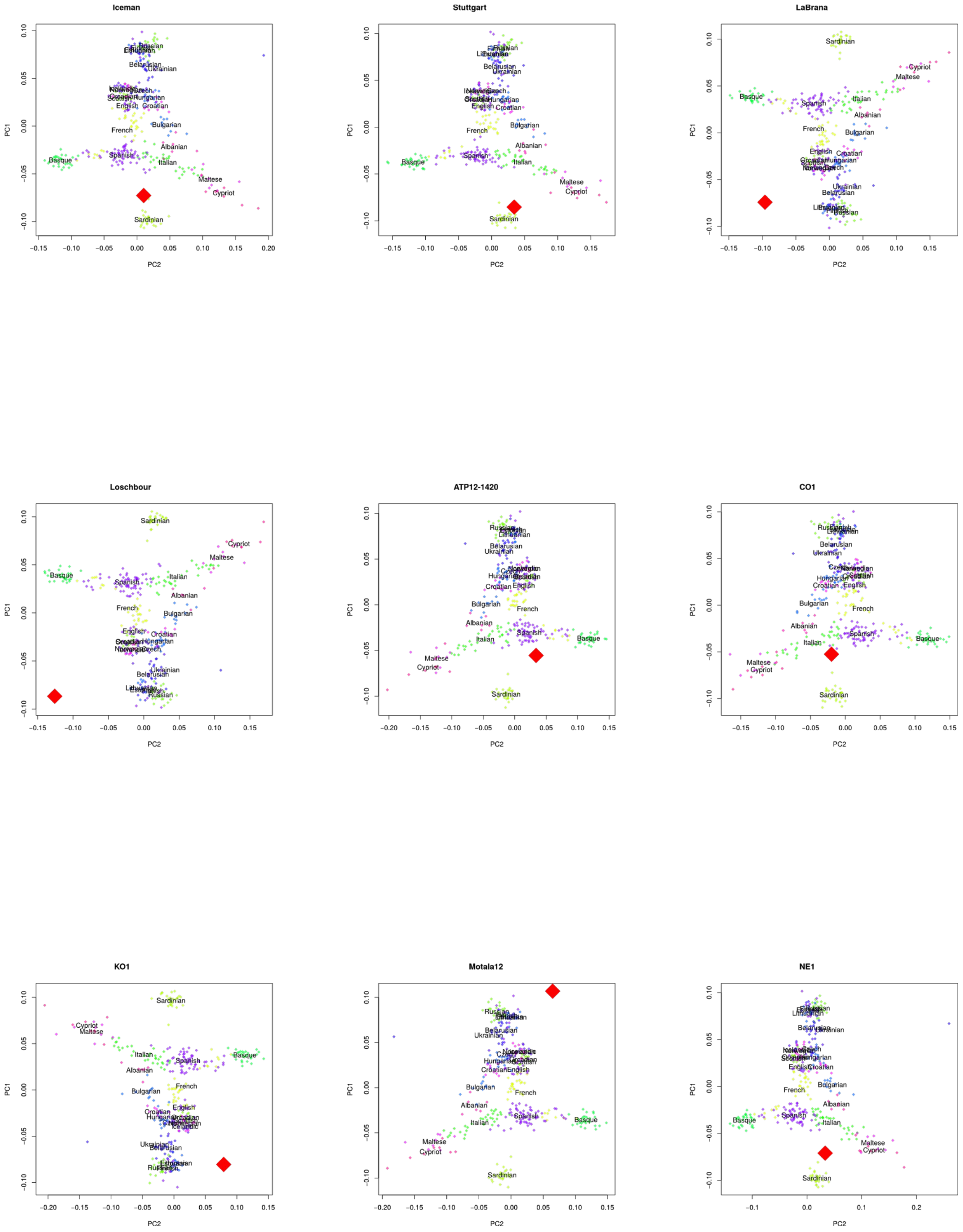


Figure S5 (continued)



Iceman

Stuttgart

LaBrana

Loschbour

ATP12-1420

CO1

KO1

Motala12

NE1

Figure S5 (continued)

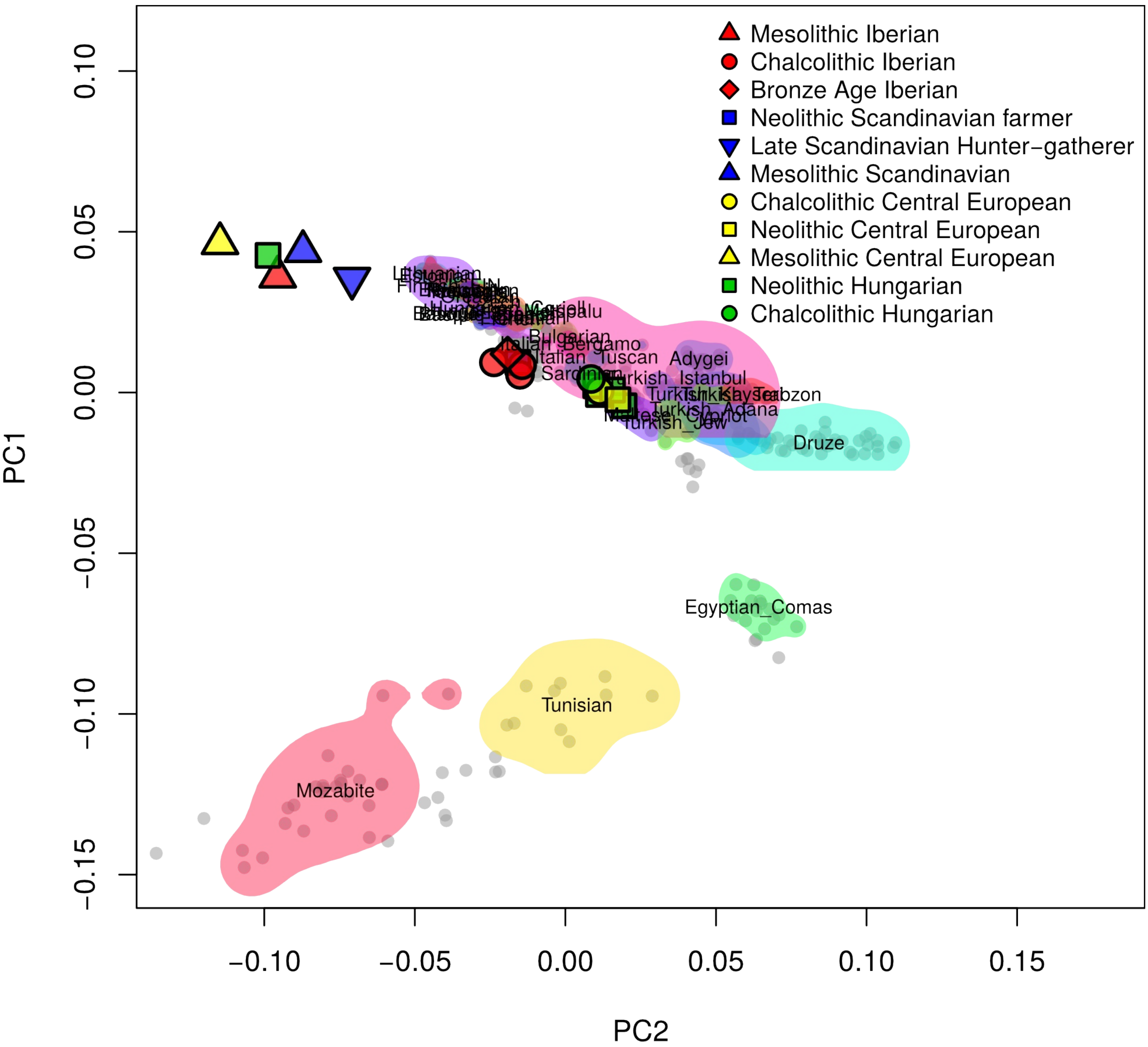


Figure S6: PCA of all ancient samples together with Western Eurasian and North African populations.

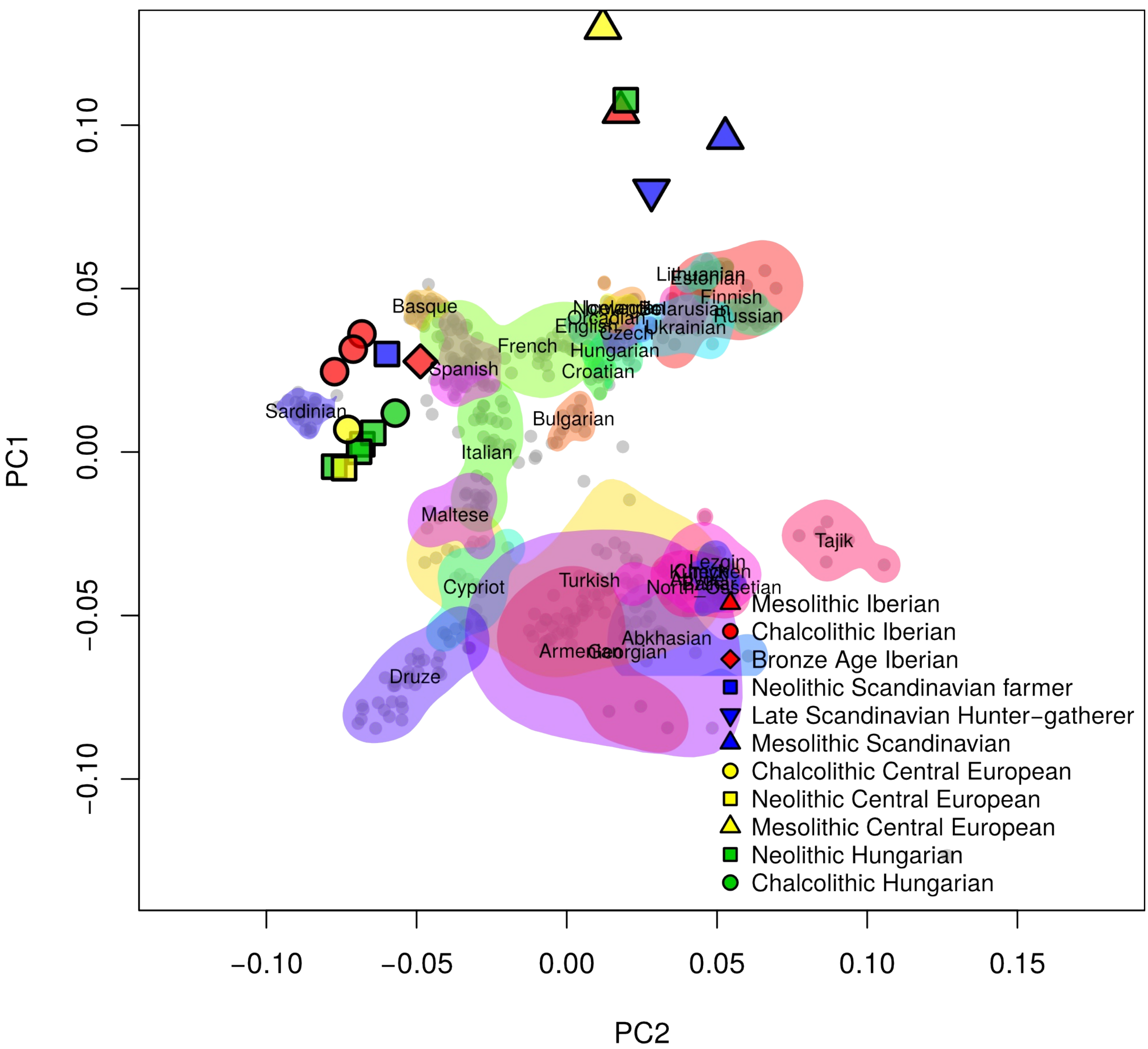


Figure S7: PCA of all ancient samples together with Western Eurasian and Caucasian populations.

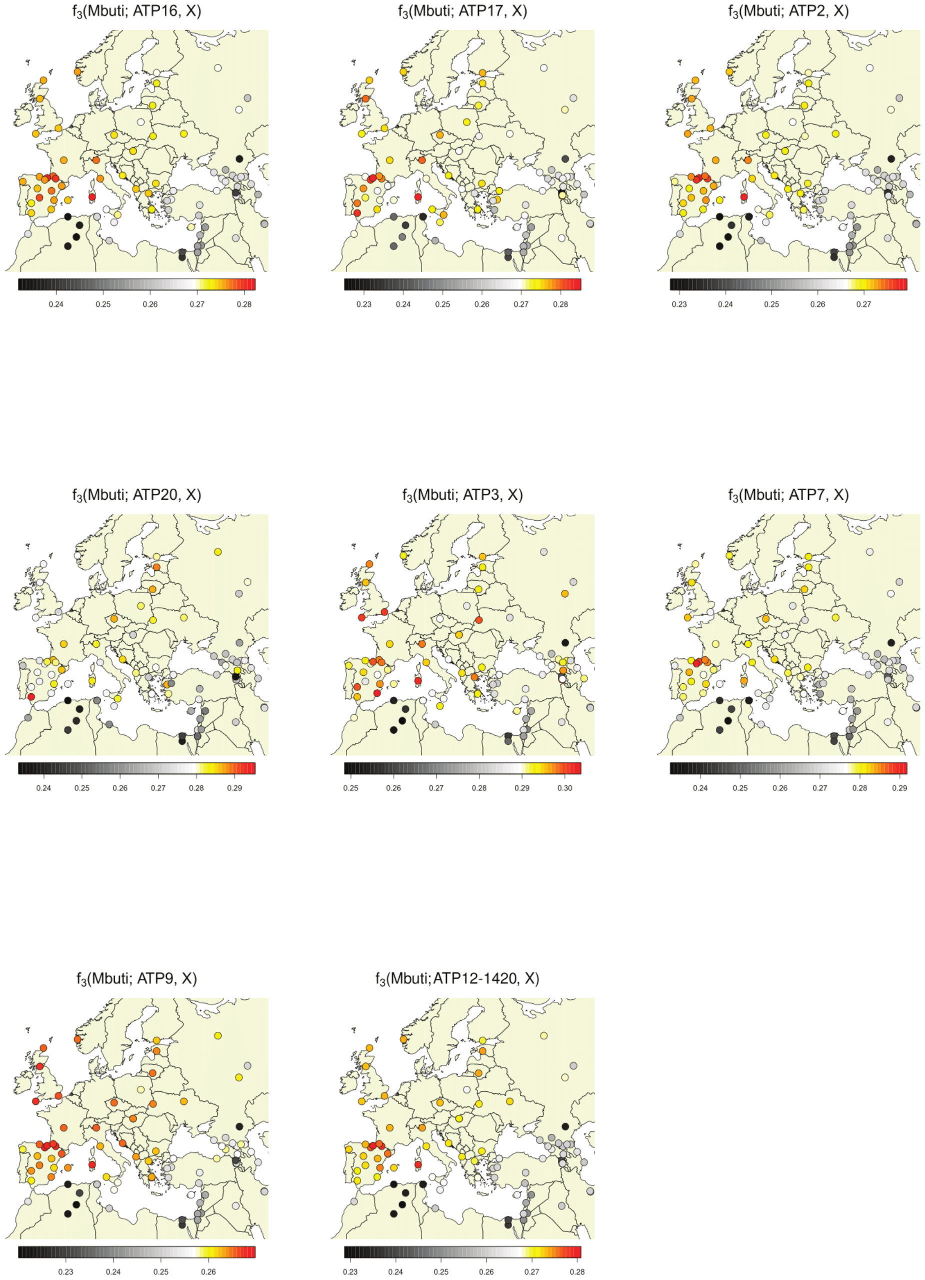


Figure S8: Outgroup f3 statistics (measuring shared drift between pairs of populations) for all eight El Portalón individuals and Western Eurasian populations from the Human Origins data set. All ancient farmers share most drift with present-day South-Western Europeans.

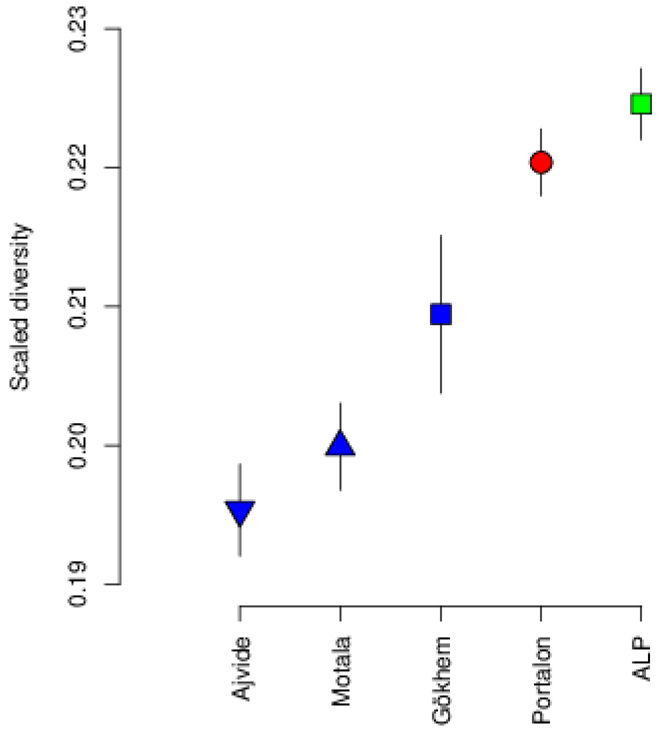


Figure S9: Scaled diversity of different ancient populations.

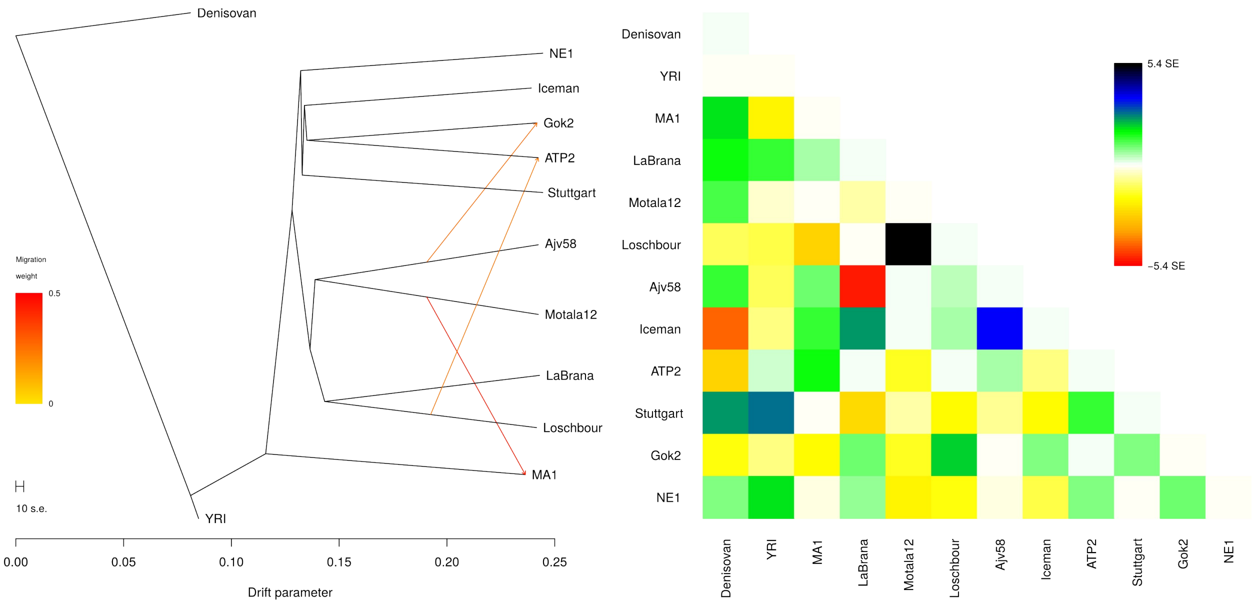


Figure S10: Treemix results with four migration events.

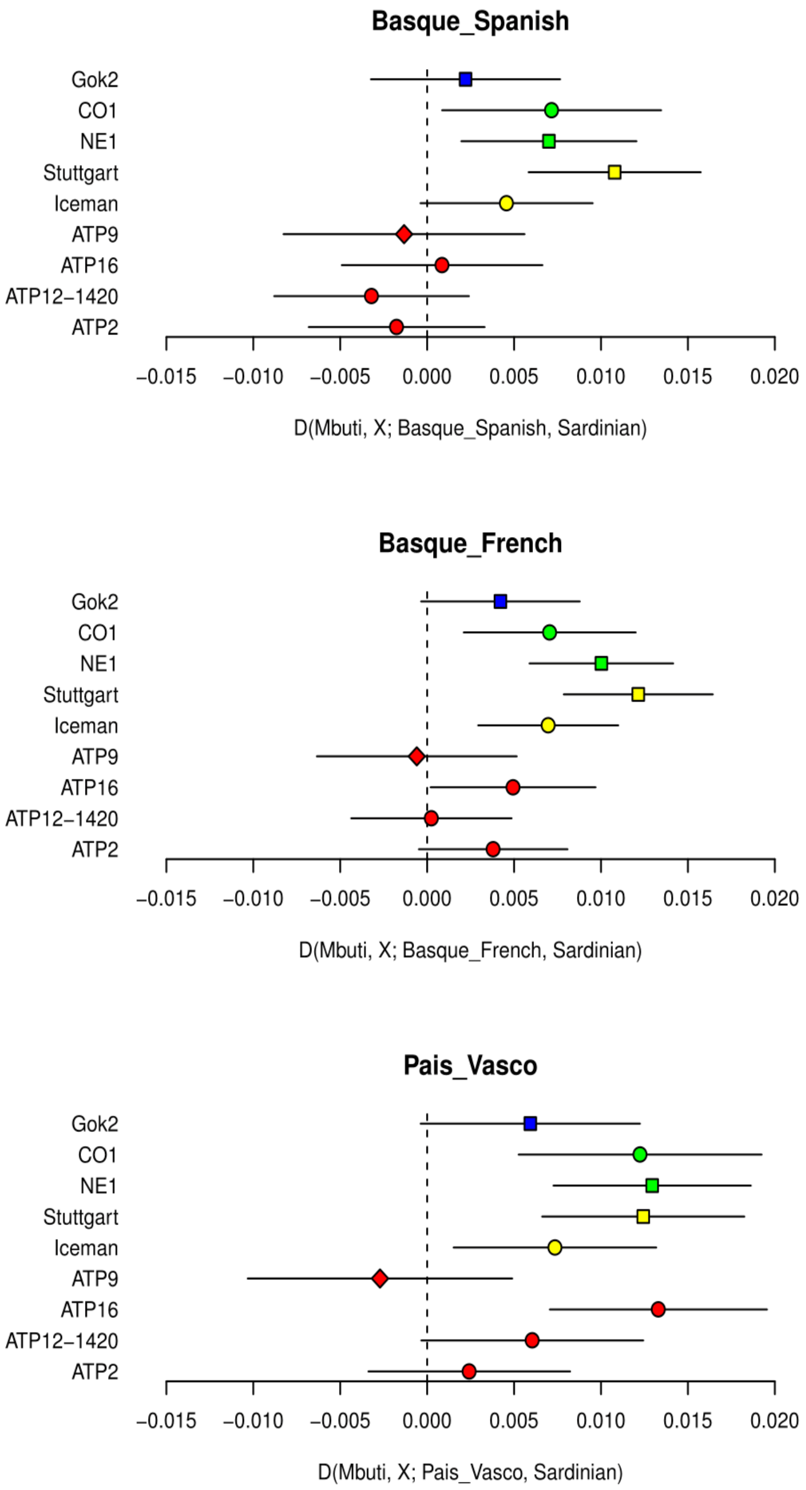


Figure S11: Genomic affinities of eight ancient farmers with Sardinians and three different Basque populations. Error bars show two standard errors.

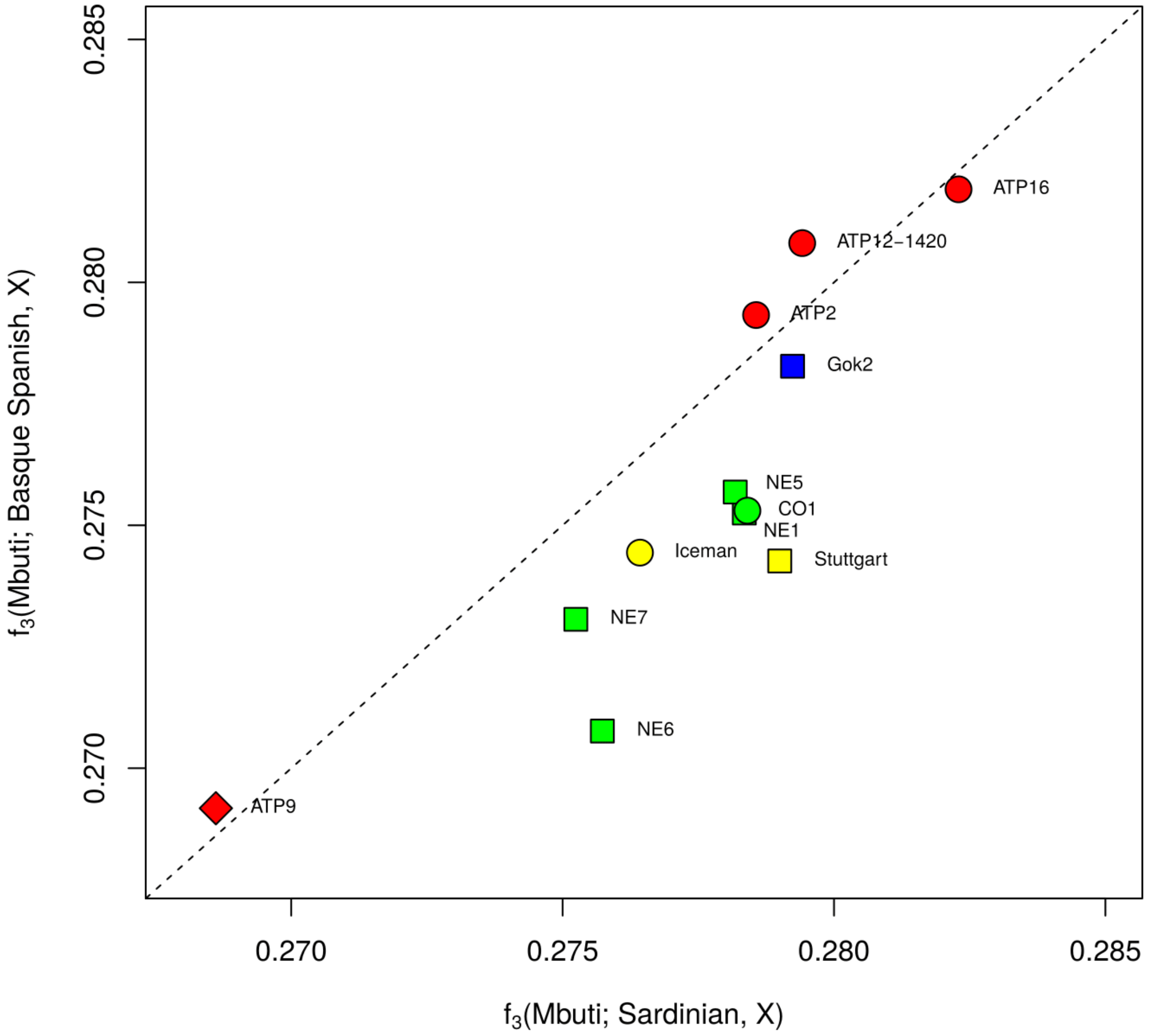


Figure S12: Shared drift measured as outgroup f3 for all ancient farmers with Sardinians and Spanish Basques. The El Portalón individuals and Gok2 share similar amounts of drift with both modern populations whereas all other farmers share more drift with Sardinians.

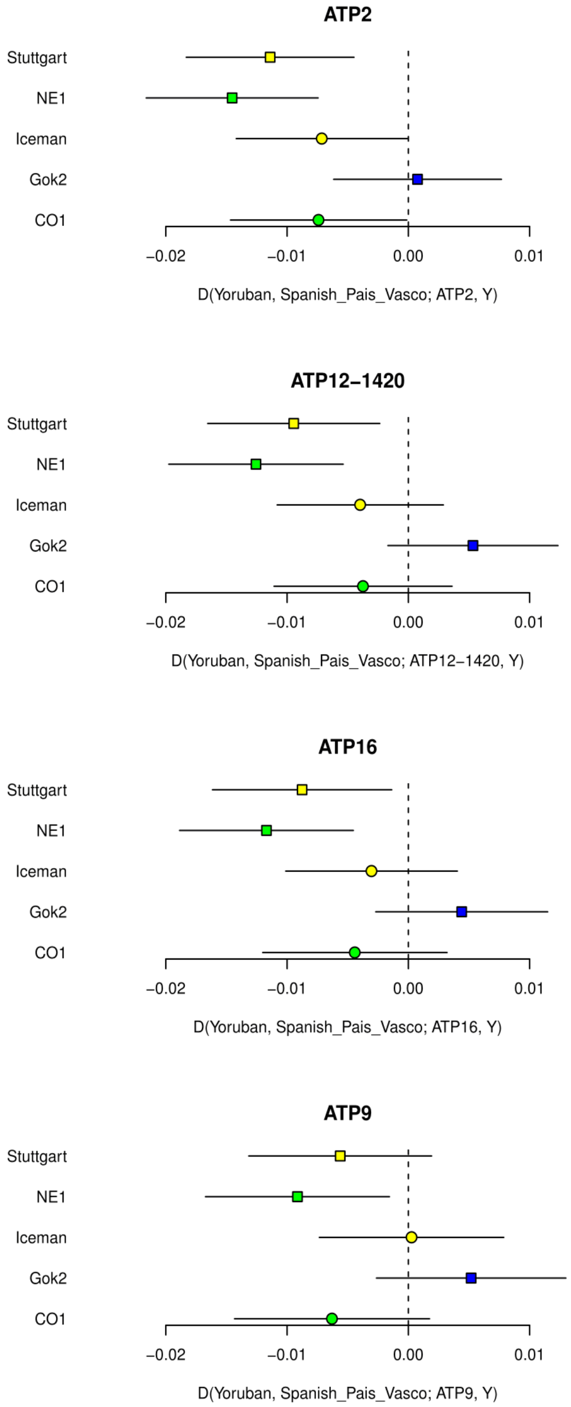


Figure S13: Genomic affinities of modern-day Basques to different ancient farmers. Negative values imply stronger affinities with El Portalón individuals. Error bars show two standard errors.

Table S1: Dating and archaeological context of the sequenced samples.

Sample ID & Provenance	aDNA lab code	Radiocarbon lab code	Conventional Radiocarbon Age (yr BP)	Age (cal BP)	Cultural Context	Description
Clandestine pit						
ATP-HUMAN 2 †	ATP2	Beta-386394	4210±30	4849-4628	Pre-bell Beaker/Chalcolithic	Almost complete subadult right femur
ATP-HUMAN 3 †	ATP3	Beta-368281	4650±30	5466-5312	Pre-bell Beaker/Chalcolithic	Adult right humeral shaft
Level 7/8 - Funerary Tumulus						
ATP-HUMAN 7	ATP7	Beta-368285	4460±40	5295-4894	Pre-bell Beaker/Chalcolithic	Fragment of a right tibial shaft
ATP-HUMAN 16	ATP16	Beta-368289	4400±30	5211-4866	Pre-bell Beaker/Chalcolithic	Adult torathic vertebra
ATP-HUMAN 17	ATP17	Beta-368290	4280±30	4957-4821	Pre-bell Beaker/Chalcolithic	Adult proximal foot phalange
ATP-HUMAN 20	ATP20	Beta-368293	3770±30	4239-4000	Pre-bell Beaker/Chalcolithic	Adult right metacarpal 3
ATP-HUMAN 21 (ATP12'-1420)	MATOJO	Beta-368295	4300±30	4960-4829	Pre-bell Beaker/Chalcolithic	Complete subadult skeleton
Main Gallery						
ATP-HUMAN 9	ATP9	Beta-386395	3390±30	3700-3568	Middle Bronze Age	Complete adult right calcaneus

(†) Samples from disturbed sediments but associated to Chalcolithic cultural context based on their absolute chronology.

Table S2: Summary of obtained information, genetic and radiocarbon, from the 8 El Portalón individuals in this study.

Individual	Percent human DNA[§]	Genome coverage	mt coverage	mt haplogroup	Y haplogroup	Biological sex	Age (C14 cal BP)	Contamination [25]	95% CI	Informative sites	Consensus alleles	Alternative alleles	proportion authentic [26]	Average fragment length[¶] [bp]
ATP2	39.1-40.9	4.08	341.2	U5b3	H2	XY	4849-4628	0.1	0-0.3	4	1444	2	0.983-0.999	61.2
ATP3	1.9-2.5	0.03	14.4	K1a2b	-	XY	5466-5312	0	0-7.2	3	40	0	0.528-0.997	79.3
ATP7	0.8-2.2	0.04	16.8	J1c1b1	-	XX	5295-4894	9.2	3.8-14.6	7	99	10	0.663-0.997	79.8
ATP9	4.8-23.8	0.41	35.4	U5b1b	-	XX	3700-3568	1.4	0-3.3	4	144	2	0.836-0.999	84.6
ATP16	5.4-9.9	1.11	68.7	X2c	-	XX	5211-4866	1.4	0.2-2.7	5	342	5	0.906-0.999	86.2
ATP17	1.9-2.0	0.03	3.6	H3	-	XY	4957-4821	-	-	-	-	-	0.080-0.990	99
ATP20	1.6-1.7	0.01	3.6	U5a1c	-	XX	4239-4000	-	-	-	-	-	0.083-0.990	79.5
ATP12-1420	6.0-8.2	1.21	230.8	H3c	I2a2a	XY	4960-4829	0.5	0-1.1	2	582	3	0.971-0.999	81.5

[§] varying between different libraries

[¶] only sequences longer than 35bp and mapping to the human genome with a minimum mapping score of 30 were considered

Table S3: X chromosome based estimates of contamination. All estimates are in percent.

Individual	**Method 1**		**Method2**	
	Point estimate	**Standard error**	**Point estimate**	**Standard error**
ATP12-1420	1.675	0.161	0.788	0.149
ATP2	2.758	0.152	2.553	0.227
Loschbour	0.778	0.044	0.837	0.111
Iceman	1.483	0.085	1.270	0.141
LaBrana1	1.041	0.099	0.825	0.129
KO1	0.454	0.090	0.176	0.109
NE5	0.888	0.181	0.495	0.123
NE6	0.456	0.125	0.270	0.112
NE7	0.803	0.138	0.452	0.113
Motala12	0.221	0.053	0.109	0.039
MA1	0.614	0.098	0.473	0.122
Ajv58	1.173	0.128	0.918	0.136

Table S4: Mitochondrial haplogroup defining markers.

Individual	Haplogroup	SNPs defining called hg	Other diagnostic SNPs
ATP2	U5b3	146T, 150T, 152T, 228A, 247G, 769G, 825T, 1018G, 2758G, 2885T, 3197C, 3594C, 4104A, 4312C, 7146A, 7226A, 7256C, 7521G, 8468C, 8655C, 8701A, 9477A, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11467G, 11914G, 12308G, 12372A, 12705C, 13105A, 13276A, 13506C, 13617C, 13650C, 14182C, 16129G, 16187C, 16189T, 16192T, 16223C, 16230A, 16270T, 16278C, 16304C, 16311T	6079T, 8860A, 16519T
ATP3	K1a2b	146T, 152T, 195T, 247G, 497T, 769G, 825T, 1018G, 1811G, 2758G, 2885T, 3480G, 3594C, 4104A, 4312C, 7146A, 7256C, 8005C, 8468C, 8655C, 8701A, 9055A, 9540T, 9698C, 10550G, 10664C, 10810T, 10873T, 10915T, 11025C, 11299C, 11467G, 11914G, 12308G, 12372A, 12705C, 13105A, 13276A, 13506C, 13650C, 14167T, 16093C, 16129G, 16187C, 16189T, 16223C, 16224C, 16230A, 16278C	12397G
ATP7	J1c1b1	146T, 152T, 185A, 195T, 228A, 247G, 295T, 482C, 769G, 825T, 1018G, 2758G, 2885T, 3010A, 3394C, 3594C, 4104A, 4312C, 5773A, 7146A, 7184G, 7256C, 7521G, 8468C, 8655C, 8701A, 9540T, 10688G, 10810T, 10873T, 10915T, 11914G, 12705C, 13105A, 13276A, 13506C, 13650C, 14798C, 16069T, 16126C, 16129G, 16187C, 16189T, 16223C, 16230A, 16278C, 16311T	4769A, 6156T, 6169T, 6172T, 6173T, 16519T
ATP9	U5b1b	146T, 150T, 152T, 195T, 247G, 769G, 825T, 1018G, 2758G, 2885T, 3197C, 3594C, 4104A, 4312C, 5656G, 7146A, 7256C, 7521G, 7768G, 8468C, 8655C, 8701A, 9477A, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11467G, 11914G, 12308G, 12372A, 12705C, 13105A, 13276A, 13506C, 13617C, 13650C, 14182C, 16129G, 16187C, 16192T, 16223C, 16230A, 16270T, 16278C, 16311T	8838A, 10172A, 11878C, 12630A, 13722G, 16221T, 16362C, 16519T
ATP16	X2c	146T, 152T, 153G, 225A, 227G, 247G, 769G, 825T, 1018G, 1719A, 2758G, 2885T, 3594C, 4104A, 4312C, 6371T, 7146A, 7256C, 7521G, 8468C, 8655C, 8701A, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11914G, 13105A, 13276A, 13506C, 13650C, 13966G, 14470C, 16129G, 16187C, 16230A, 16255A, 16311T	15314A, 16207G
ATP17	H3	73A, 146T, 152T, 195T, 247G, 769G, 825T, 1018G, 2706A, 2758G, 2885T, 3594C, 4104A, 4312C, 6776C, 7256C, 8468C, 8655C, 8701A, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11719G, 11914G, 12705C, 13105A, 13276A, 13506C, 13650C, 14766C, 16129G, 16187C, 16189T, 16223C, 16230A, 16278C, 16311T	4769A, 7702A, 7729G, 8573A, 9182A, 16093C
ATP20	U5a1c1a	146T, 152T, 153G, 247G, 769G, 825T, 1018G, 2758G, 2885T, 3197C, 3594C, 4312C, 7256C, 8468C, 8655C, 8701A, 9477A, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11467G, 11914G, 12308G, 12372A, 12705C, 13105A, 13276A, 13506C, 13617C, 13650C, 13802T, 16129G, 16187C, 16189T, 16192T, 16223C, 16230A, 16256T, 16278C, 16320T, 16399G	195T, 7013A, 14766C, 16519T
ATP12-1420	H3c	73A, 146T, 152T, 195T, 247G, 769G, 825T, 1018G, 2706A, 2758G, 2885T, 3594C, 4104A, 4312C, 6776C, 7028C, 7146A, 7256C, 7521G, 8468C, 8655C, 8701A, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11719G, 11914G, 12705C, 12957C, 13105A, 13276A, 13506C, 13650C, 14766C, 16129G, 16187C, 16189T, 16223C, 16230A, 16278C, 16311T	7493T, 8860A, 16399G

Table S5: Y-Haplogroup defining markers for ATP2.

Position (hg19)	Marker-ID	Mutation	Allele in ATP2	Associated Haplogroup	Sequencing Depth
7374927	L985	A->C	C	A1	2
18759708	L1053	C->A	A	A1	2
21646577	L1013	C->A	A	A1	4
9108252	P187	G->T	T	F	4
14832620	M235	T->G	G	F	1
18097251	P159	C->A	A	F	1
6932824	L279	G->T	T	H2	3
8353840	L281	T->G	G	H2	3
14869743	P96	C->A	A	H2	2

Table S6: Y-Haplogroup defining markers for ATP12-1420.

Position (hg19)	Marker-ID	Mutation	Allele in ATP12-1420	Associated Haplogroup	Sequencing Depth
18759708	L1053	C->A	A	A1	1
19038302	P124	A->C	C	IJ	1
7879415	L460	A->C	C	I2a	1
7716262	S151, L34	A->C	C	I2a2a	2
8353707	P221, S120	C->A	A	I2a2a	1

Table S7: Allelic state at major SNPs involved in appearance and lactase persistence.

SNP	Gene	Phenotype	ATP2	ATP3	ATP7	ATP9	ATP16	ATP17	ATP20	ATP12-1420	La Brana	Ref
rs4988235	LCT	Lactase persistence	A(9)				A(3)			A(1)	A(6)/D*(1)	[123]
rs1426654	SLC24A5	Pigmentation	D(7)			D(1)				D(3)	A(3)	[124]
rs16891982	SLC45A2	Pigmentation	A(7)			A(1)/D(1)	A(2)/D(1)			A(1)	A(6)	[125]
rs1805007	MC1R	Red hair/fair skin	A(5)/D*(1)			A(2)	A(3)				A(7)	[126]
rs12913832	HERC2/ OCA2	Brown/blue eyes	A*(2)/D(3)				A(3)			A(5)	D(3)	[127]

A='ancestral', D='derived', *=potential post mortem damage, allele counts are given in parentheses

Table S8: Transversion SNPs overlapping with the Human Origins Array for all ancient individuals used in this study.

Individual	Number of merged SNPs	Reference
ATP16	78048	this study
ATP17	3676	this study
ATP20	1959	this study
ATP2	107245	this study
ATP3	3749	this study
ATP7	5126	this study
ATP12-1420	84616	this study
ATP9	42498	this study
Ajv58	98091	[20]
Gok2	74797	[20]
Iceman	108306	[74]
LaBrana1	105040	[73]
Stuttgart	110511	[19]
Loschbour	110644	[19]
Motala12	99558	[19]
Motala1	17791	[19]
Motala2	15434	[19]
Motala3	47941	[19]
MA1	79226	[76]
NE1	110168	[71]
NE2	15891	[71]
NE5	61857	[71]
NE6	71201	[71]
NE7	68584	[71]
CO1	64895	[71]
KO1	74657	[71]

Table S9: Transversion SNPs overlapping with the SNPs polymorphic in Yorubans for ancient individuals used in sequence comparisons.

Individual	**Overlapping transversion SNPs**
CO1	927,597
ATP9	595,912
ATP12-1420	1,240,025
ATP16	1,188,376
ATP2	1,609,208
MA1	1,080,499
Motala12	1,449,550
Loschbour	1,721,454
Ajv58	1,408,191
LaBrana	1,435,379
Stuttgart	1,702,338
Gok2	880,246
Iceman	1,401,203
NE1	1,709,509

Table S10: Results for Treemix allowing for four migration events.

Source	**Target**	**Admixture proportion**	**Standard error**	**P value**
Motala12	MA1	0.45642	0.0447847	<2.22507e-308
Ajv58	Gok2	0.325296	0.0369275	<2.22507e-308
(LaBrana, Loschbour)	ATP2	0.300488	0.0300551	2.95E-010
Ajv58	Iceman	0.186735	0.036777	1.11E-016

Table S11: Results for Treemix estimating the weight of pre-defined migration events.

Source	**Target**	**Admixture proportion**	**Standard error**	**P value**
Ajv58	Gok2	0.140017	0.0316838	4.44E-006
LaBrana	Gok2	0.0698887	0.0399433	0.0382792
Loschbour	Gok2	0.107201	0.0392299	0.00295297
Ajv58	Iceman	0.106289	0.022066	7.89E-007
LaBrana	Iceman	0.0790324	0.0158708	3.69E-007
Loschbour	Iceman	0.0439413	0.0243951	0.0369624
Ajv58	ATP2	0.0769008	0.0180367	1.06E-005
LaBrana	ATP2	0.102443	0.0162427	1.50E-010
Loschbour	ATP2	0.100224	0.0302679	0.000503037
Ajv58	CO1	0.0831047	0.0154387	3.59E-008
LaBrana	CO1	0.0742148	0.017967	1.66E-005
Loschbour	CO1	0.0560112	0.0183718	0.00106859

Table S12: Sequence based Dtest results for affinities of ancient farmers to different HG populations. D statistics were calculated as D(Yorubans, ancient famer; HG1, HG2).

Farmer	HG1	HG2	D	Standard error	Z
ATP2	LaBrana	Loschbour	0.0100089055	0.005918496	1.6911231391
ATP2	Ajv58	Loschbour	0.011999459	0.006038529	1.9871493455
ATP2	Ajv58	LaBrana	0.0020727261	0.0056880863	0.364397792
ATP2	Motala12	Loschbour	0.0272963351	0.0060047387	4.545799019
ATP2	Motala12	LaBrana	0.0169569597	0.0058465701	2.9003260647
ATP2	Motala12	Ajv58	0.013187086	0.0059559941	2.2140864614
Iceman	LaBrana	Loschbour	-0.0031948685	0.0058733537	-0.5439598283
Iceman	Ajv58	Loschbour	-0.0130054155	0.0059790477	-2.175165018
Iceman	Ajv58	LaBrana	-0.0109442444	0.0055230748	-1.9815491963
Iceman	Motala12	Loschbour	0.0036374401	0.005805799	0.6265184394
Iceman	Motala12	LaBrana	0.0068744066	0.0056967946	1.206714849
Iceman	Motala12	Ajv58	0.0166183365	0.0057981467	2.8661462765
NE1	LaBrana	Loschbour	-0.0016330068	0.0057287663	-0.2850538452
NE1	Ajv58	Loschbour	-0.0074982355	0.0057387591	-1.3065952601
NE1	Ajv58	LaBrana	-0.0071108574	0.0054098946	-1.3144170001
NE1	Motala12	Loschbour	0.0052839183	0.0058906811	0.8969961539
NE1	Motala12	LaBrana	0.0062377409	0.0057035371	1.0936618502
NE1	Motala12	Ajv58	0.0115210123	0.0058517405	1.968818064
Stuttgart	LaBrana	Loschbour	0.0099853325	0.005770986	1.7302645502
Stuttgart	Ajv58	Loschbour	-0.0002055653	0.0058448342	-0.0351704305
Stuttgart	Ajv58	LaBrana	-0.0122728266	0.0055481769	-2.2120467473
Stuttgart	Motala12	Loschbour	0.0041721847	0.00577113	0.7229406817
Stuttgart	Motala12	LaBrana	-0.0057968941	0.0056313737	-1.0293925392
Stuttgart	Motala12	Ajv58	0.0045178295	0.0056456527	0.8002315644
Gok2	LaBrana	Loschbour	0.0114799548	0.0059820572	1.9190646946
Gok2	Ajv58	Loschbour	-0.0142311176	0.0059269545	-2.4010843472
Gok2	Ajv58	LaBrana	-0.0236128572	0.0058074948	-4.0659282397
Gok2	Motala12	Loschbour	0.0144749405	0.0060112922	2.4079582159
Gok2	Motala12	LaBrana	0.0023842737	0.0059563059	0.4002940332
Gok2	Motala12	Ajv58	0.029305679	0.0059710692	4.907945

Table S13: Correlation of age and f_4(Mbuti, ref_HG; ref_farmer, farmer) for different reference individuals.

ref_HG	ref_farmer	R^2(f4(Mbuti, ref_HG; ref_farmer, farmer), calBP(farmer))
Loschbour	Stuttgart	0.6167983135
Loschbour	NE1	0.5811809207
Loschbour	NE5	0.5551230349
Loschbour	NE6	0.635415306
LaBrana	Stuttgart	0.5114420387
LaBrana	NE1	0.4893954384
LaBrana	NE5	0.4374938774
LaBrana	NE6	0.6377920928
Motala12	Stuttgart	0.2965967078
Motala12	NE1	0.3025798589
Motala12	NE5	0.2252426015
Motala12	NE6	0.249633664